Pattern Recognition 167 (2025) 111715

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Self-adaptive image-text fusion for medical image classification

Jian Zhang [a], Kaihao He [b], Zunlei Feng [c], Shuifa Sun [a,*], Xiaoyan Sun [a], Zhenming Yuan [a], Jun Yu [b]

[a] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China
[b] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China
[c] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310007, China

ARTICLE INFO

Keywords:
Multimodal fusion
Image classification
Cross-attention
Feature fusion

ABSTRACT

Multimodal classification using both medical images and text reports propels the computer aided disease diagnosis. The performance is susceptible to the quality of image-text fusion. Due to the semantic gap and weak correlation between image and text, current image-text fusion approaches cannot achieve satisfactory results. We propose a self-adaptive image-text fusion approach to multimodal medical image classification. We learn a mapping from image to text to achieve semantic alignment that mitigates the inter-modality semantic gap, and estimate a binary correlation mask with Jensen–Shannon Divergence (JSD) loss to retrieve image and text features that have strong correlations to achieve feature alignment. Then, we propose a parameter-free feature fusion method based on a Simplified-Attention mechanism, which queries image features using text features and concatenates the results to achieve computationally efficient feature fusion. We fuse all the image and text features for medical image classification. Experimental results on three datasets reveal that the proposed approach outperforms a group of state-of-the-art methods, and demonstrates superior medical interpretability.

1. Introduction

Medical images and text reports have been the basis of medical diagnosis for many years. In computer aided diagnosis, the information from images and reports should be well fused to boost the diagnosis accuracy. Though multimodal image fusion [1] has been well studied, the fusion between medical images and texts [2] is far behind the need owing to the obvious heterogeneity of the two modalities.

Multimodal fusion methods can be categorized as early (data level) fusion, intermediate (feature level) fusion, and late (decision level) fusion [3]. Early fusion aims to merge the original data from different modalities [4]. Due to the structure discrepancy, early fusion cannot be applied to image-text fusion. Late fusion makes final decision by leveraging the decisions derived from each individual modality [5]. The inter-modality interaction is insufficient, which weakens the learning of complementary information between modalities. The feature-level intermediate fusion overcomes the structural discrepancy problem and ensures sufficient interaction between modalities.

We summarize the intermediate fusion methods into four types, i.e. direct fusion (feature concatenation [6], summation [7]), feature mapping (bilinear pooling [8]), model-based fusion (Long Short-Term Memory (LSTM) network [9], attention [10], transformer [11]), and contrastive learning [12]. Direct fusion is commonly used method, which also acts as ingredients of model-based fusion methods. Some base models, e.g. LLaVA [13], comprehensively adopt multiple types of intermediate fusion methods to achieve multimodal understanding.

However, current intermediate image-text fusion methods still suffer from three dilemas. The first one is inter-modal semantic gap. Texts naturally have strong and straightforward semantic but the image semantic can be obscure and person-dependent. The average entropy of images is much higher than that of text reports, indicating that images have much more complex semantic than texts. Aligning text semantic to image introduces extra information into text features that could be uncertain and harmful to feature fusion. Second, the correlation between image and text is weak and dynamic. The weak correlation means not all words are exactly related to image content while some image appearances may have no lexical descriptions. The dynamic correlation means that the correlated image and text content varies in different patients. Therefore, forcing the two modalities to directly align with each other may degrade the effect of multimodal feature fusion. Third, the fusion methods can be computationally complex. Model-based fusion methods rely on complex feature transformations defined by extra parameters. Contrastive learning needs to compute

* Corresponding author.
E-mail address: watersun@hznu.edu.cn (S. Sun).

https://doi.org/10.1016/j.patcog.2025.111715
Received 22 December 2024; Received in revised form 15 April 2025; Accepted 16 April 2025
Available online 2 May 2025

pairwise similarities between each sample and all other samples in loss function.

To this end, this paper proposes a self-adaptive image-text fusion approach to medical image classification. This approach consists of an Interpretable Multimodal Alignment (IMA) module and a Simplified-Attention-based Feature Fusion (SAFF) module. In IMA module, we use a learnable cross-modal mapping to align the semantic of image features to that of the text features, then a binary mask is estimated from the image and text features and applied to both modalities to adaptively select pairwise features with strong inter-modality correlations. We then exert Jensen–Shannon Divergence (JSD) loss function on the selected features to achieve feature-level alignment between the two modalities. In SAFF module, we remove the parameters of attention mechanism, and exploit a cross-attention method to fuse the aligned image and text features. Based on this image-text fusion approach, we conduct multimodal medial image classification. Experimental results on three datasets show that the proposed approach outperforms unimodal classification and a group of state-of-the-art multimodal classification algorithms. In addition, the IMA module pays more attention to the image region that is closely related to the disease, therefore has medical interpretability.

The contribution of this approach is three-fold: (1) Cross-Modal Semantic Alignment with Adaptive Mapping: A learnable mapping aligns image features to text semantics while preserving diagnostically critical visual information. This mitigates modality heterogeneity and enhances semantic interaction between the two modalities. (2) Interpretable Feature Alignment via Dynamic Masking: A trainable binary mask filters irrelated features with the guide of a JSD loss. This enhances fusion effect and interpretability by focusing on disease-relevant regions. (3) Lightweight Fusion via Simplified-Attention: A non-parametric cross-attention strategy is applied to feature fusion to reduce training complexity while maintaining competitive accuracy on small medical datasets.

The rest part of this paper is organized as follows: Section 2 introduces the related works. Section 3 presents our approach in detail. Section 4 provides the experimental results and discussions. Section 5 concludes the paper.

2. Related works

2.1. Direct fusion

Direct fusion refers to fusing multimodal data or features through concatenation, element-wise multiplication, and summation. In [14], the model conducts element-wise multiplication of Magnetic Resonance (MR) images for brain tumor segmentation. Lin et al. [15] computed element-wise averaging of X-ray image features and text features to predict ICU mortality. Venugopalan et al. [16] concatenated the features of MR images, Single Nucleotide Polymorphisms (SNP) data and Electronic Health Records (EHR) to predict the Alzheimer's disease stage. Guarrasi et al. concatenated X-ray image features and tabular features for explainable COVID-19 stratification [17]. Li et al. [7] exploited weighted summation for multimodal medical image fusion. Amsterdam et al. [18] concatenated video and kinematic features for gesture recognition. Direct fusion methods are often superposed with one another [19] for performance improvement.

The combination of multiple direct fusion methods can be more effective, but seeking for a good combination may be a process of trial and error. In addition, direct fusion does not intentionally reduce inter-modality semantic gap. This can be well addressed by the learnable mapping and masking in our approach.

2.2. Feature mapping

Feature mapping aims to project the features of different modalities into a semantically consistent space to achieve feature fusion. The typical work is Multi-modal Factorized Bilinear (MFB) pooling [8] proposed for Visual Question Answering (VQA). Liu et al. [20] extracted image and text features through self-attention, and conducted multimodal fusion using MFB. Wei et al. [21] enhanced MFB with inter-modality correlation learning to fuse dermoscopic and clinical features for skin lesion classification. LLaVA [13] and FROMAGe [22] also exploit linear projection to map the visual embeddings to text embeddings, which are used to construct multimodal instructions to fine-tune the LLM.

Feature mapping is applied to all the features including those with weak inter-modality correlation. As comparison, we equip the mapping with binary mask to exclude irrelated features to improve feature alignment.

2.3. Model-based fusion

Multimodal fusion can be done by parametric models. A two-layer LSTM network was proposed in [9] for image-text feature fusion. Mei et al. [23] used cross-attention to fuse radiology reports and X-ray images for medical report generation. Similar methods have been adopted for brain disease diagnosis [24] and fake news detection on social networks [25]. The state-of-the-art model based on attention is transformer. Wang and Huang exploited transformers for radiology report generation [11] and medical VQA [26] respectively.

Complex parameter estimation increases the workload of training. In addition, the effect of parameters is only defined by task objective. In contrast, the only set of parameters in our method is the feature mapping, whose effect is clearly defined by the JSD loss. So our approach is simpler and more interpretable.

2.4. Fusion via contrastive learning

Contrastive learning can be used for semantic alignment of multimodal data. CLIP [27] aligns image and text features by maximizing the similarity between image and text with same semantic while minimizing the similarity between those with different semantic. Hou et al. [12] conducted self-supervised multimodal learning based on contrastive learning. However, contrastive learning needs large amount of training samples that are probably unavailable in medical tasks, and the training brings about extra computational burden. As comparison, the feature mapping and binary mask in our approach can be computed efficiently.

3. The proposed method

3.1. Overview

We aim to learn a self-adaptive model to fuse medical images and text reports for disease classification. For given N sample images and corresponding reports, we use a visual encoder and a text encoder to extract features from images and reports respectively. Let $\{\boldsymbol{\mathcal{V}}^{(1)},\dots,\boldsymbol{\mathcal{V}}^{(N)}\}$ be the feature maps of N images. Let $\{\mathbf{E}^{(1)},\dots,\mathbf{E}^{(N)}\}$ be the token features of N reports. The framework of our approach can be depicted by Fig. 1, where the Self-Adaptive Fusion (SAF) module transforms image and text features to unified feature representations $\mathbf{Z}^{(n)}$ that are sent to a classifier to make prediction. The SAF module consists of an Interpretable Multimodal Alignment (IMA) module and a Simplified-Attention-based Feature Fusion (SAFF) module. The IMA module first learns a channel-wise shared linear mapping $\mathbf{W}$ to conduct semantic alignment of image to text. Then, IMA applies a binary mask to both modalities to select some pairwise features on which a symmetric Kullback–Leibler Divergence (KLD), i.e. Jensen–Shannon Divergence (JSD) [28] loss is exerted to achieve feature-level alignment. The SAFF module accepts all the aligned image and texture features and generates fused features for classification. The technical details will be shown in the following sections.

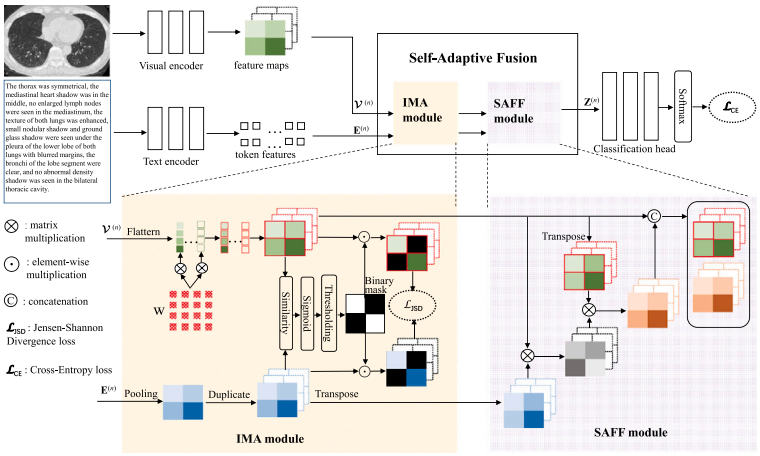

Fig. 1. The framework of our approach. The IMA module learns mapping $\mathbf{W}$ to achieve semantic alignment of image to text, then derives a binary mask from two modalities for feature selection and correlation. SAFF module fuses the aligned features for subsequent classification.

3.2. Multimodal feature extraction

We adopt a VGG-11 network as the visual encoder. Let $\mathcal{V}^{(n)} \in \mathbb{R}^{H\times W\times C}$ be the n_{th} image's feature maps in the last convolution layer, where H, W and C denotes the height, width of a feature map and the number of feature maps (channels). We use $\mathbf{V}_c^{(n)} \in \mathbb{R}^{H\times W}$ to denote the c_{th} feature map in tensor $\mathcal{V}^{(n)}$, use $\mathbf{v}_{hw:}^{(n)} \in \mathbb{R}^{C\times 1}$ to denote the features at the h_{th} row and w_{th} column of all the channels. Vector $\mathbf{v}_{:wc}^{(n)} \in \mathbb{R}^{H\times 1}$ and $\mathbf{v}_{h:c}^{(n)} \in \mathbb{R}^{1\times W}$ denotes a column and a row in the c_{th} feature map.

The text report is tokenized and sent to cascaded two transformers to yield the features. Let the features of the n_{th} report be $\mathbf{E}^{(n)} \in \mathbb{R}^{T\times L}$, where T is the token dimension and L is the token number. We divide $\mathbf{E}^{(n)}$ into $H \times W$ blocks to guarantee each block is approximately a $\frac{T}{H} \times \frac{L}{W}$ submatrix of $\mathbf{E}^{(n)}$. Then we choose the max value within each block to represent $\mathbf{E}^{(n)}$, and duplicate the pooled $\mathbf{E}^{(n)}$ for C times to obtain the text tensor $\mathcal{U}^{(n)} \in \mathbb{R}^{H\times W\times C}$. Similarly, we use $\mathbf{U}_c^{(n)} \in R^{H\times W}$, $\mathbf{u}_{hw:}^{(n)} \in \mathbb{R}^{C\times 1}$, $\mathbf{u}_{:wc}^{(n)} \in \mathbb{R}^{H\times 1}$, and $\mathbf{u}_{h:c}^{(n)} \in \mathbb{R}^{1\times W}$ to denote the ingredients of $\mathcal{U}^{(n)}$.

3.3. Self-adaptive multimodal fusion

3.3.1. Interpretable multimodal alignment

This module learns a semantic mapping from the image feature space to text feature space. We improve the channel-wise fully-connected layer [29] by letting all the channels share the same connection weights (these channels reflect the same semantic of an image). To this end, we flatten each feature map $\mathbf{V}_c^{(n)}$ into a $(H \times W) \times 1$ vector by stacking each row's (of $\mathbf{V}_c^{(n)}$) transposition, then multiply this vector with a learnable semantic mapping matrix $\mathbf{W}$ with size $(H \times W)\times(H \times W)$ to achieve semantic alignment. We reshape the matrix–vector product to size $H\times W$ and obtain the semantically aligned image feature map $\tilde{\mathbf{V}}_c^{(n)}$. The aligned image tensor can be denoted by $\tilde{\mathcal{V}}^{(n)}$. This operation can be depicted as (1):

$$\begin{pmatrix} \left(\tilde{\mathbf{v}}_{1:c}^{(n)}\right)^T \\ \vdots \\ \left(\tilde{\mathbf{v}}_{H:c}^{(n)}\right)^T \end{pmatrix} = \mathbf{W}\cdot\begin{pmatrix} \left(\mathbf{v}_{1:c}^{(n)}\right)^T \\ \vdots \\ \left(\mathbf{v}_{H:c}^{(n)}\right)^T \end{pmatrix}, \qquad \tilde{\mathbf{V}}_c^{(n)} = \begin{pmatrix} \tilde{\mathbf{v}}_{1:c}^{(n)} \\ \vdots \\ \tilde{\mathbf{v}}_{H:c}^{(n)} \end{pmatrix}, \tag{1}$$

where T is transpose. Fig. 2 shows an example of the interpretable multimodal alignment procedure, where the related computation pipeline is rendered with red arrows. The data in Fig. 2 are actually estimated values with precision rounding by neural network, in which we set the size of $\mathcal{V}^{(n)}$ as $2 \times 2 \times 2$.

After semantic alignment, we can compute a correlation mask $\mathbf{M}^{corr} \in \mathbb{R}^{H\times W}$ based on $\tilde{\mathcal{V}}^{(n)}$ and $\mathcal{U}^{(n)}$. The element of $\mathbf{M}^{corr}$ at the h_{th} row and w_{th} column represents the cosine similarity between $\mathbf{u}_{hw:}^{(n)}$ and $\tilde{\mathbf{v}}_{hw:}^{(n)}$ with Sigmoid activation. We then leverage a thresholding strategy to obtain a binary mask $\mathbf{M}^{bin}$ where 1 means the activation value is larger than a predefined threshold while 0 means not. The computation can be formulated as:

$$m_{hw}^{corr} = \mathrm{Sig}\left(\frac{\left(\mathbf{u}_{hw:}^{(n)}\right)^T \cdot \tilde{\mathbf{v}}_{hw:}^{(n)}}{\|\mathbf{u}_{hw:}^{(n)}\| \cdot \|\tilde{\mathbf{v}}_{hw:}^{(n)}\|}\right), \quad \mathbf{M}^{bin} = \mathrm{Bin}\left(\mathbf{M}^{corr}\right), \tag{2}$$

where m_{hw}^{corr} represents the element of $\mathbf{M}^{corr}$ at the h_{th} row and w_{th} column, $\mathrm{Sig}(\cdot)$ denotes Sigmoid activation and $\mathrm{Bin}(\cdot)$ represents the thresholding operation. We multiply $\mathbf{M}^{bin}$ with $\mathbf{U}_c^{(n)}$ and $\tilde{\mathbf{V}}_c^{(n)}$ element by element to select strongly correlated features $\tilde{\mathbf{V}}_c^{(n)}$ and $\tilde{\mathbf{U}}_c^{(n)}$ from both modalities: $\tilde{\mathbf{V}}_c^{(n)} = \mathbf{M}^{bin} \odot \tilde{\mathbf{V}}_c^{(n)}$, $\tilde{\mathbf{U}}_c^{(n)} = \mathbf{M}^{bin} \odot \mathbf{U}_c^{(n)}$, where $\odot$ denotes the element-wise multiplication.

We flatten $\tilde{\mathbf{V}}_c^{(n)}$ and $\tilde{\mathbf{U}}_c^{(n)}$ (with zero elements) into vectorized form $\vec{\mathbf{v}}_c^{(n)}$ and $\vec{\mathbf{u}}_c^{(n)}$ using the same method of stacking the transposition of each row. By further concatenating $\vec{\mathbf{v}}_c^{(n)}$ and $\vec{\mathbf{u}}_c^{(n)}$ of all channels, we derive the vectorized features of the n_{th} image and the n_{th} report, denoting as $\vec{\mathbf{v}}^{(n)}$ and $\vec{\mathbf{u}}^{(n)}$ respectively. We enhance their probabilistic similarity to guarantee the strong inter-modality correlation by minimizing the Kullback–Leibler Divergence (KLD) between $\vec{\mathbf{v}}^{(n)}$ and $\vec{\mathbf{u}}^{(n)}$:

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(\vec{\mathbf{v}}^{(n)}|\vec{\mathbf{u}}^{(n)}) &= \sum_{d=1}^{D} \vec{\mathbf{v}}^{(n)}(d)\log\left(\frac{\vec{\mathbf{v}}^{(n)}(d)}{\vec{\mathbf{u}}^{(n)}(d)}\right), \\ \mathrm{D}_{\mathrm{KL}}(\vec{\mathbf{u}}^{(n)}|\vec{\mathbf{v}}^{(n)}) &= \sum_{d=1}^{D} \vec{\mathbf{u}}^{(n)}(d)\log\left(\frac{\vec{\mathbf{u}}^{(n)}(d)}{\vec{\mathbf{v}}^{(n)}(d)}\right), \end{aligned} \tag{3}$$

where d means the d_{th} vector element and D is the feature dimension. Accordingly, the image-text feature alignment (illustrated with black arrows in Fig. 2) can be realized by minimizing the Jensen–Shannon Divergence (JSD) loss:

$$\mathcal{L}_{JSD} = \sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}(\vec{\mathbf{v}}^{(n)}|\vec{\mathbf{u}}^{(n)}) + \mathrm{D}_{\mathrm{KL}}(\vec{\mathbf{u}}^{(n)}|\vec{\mathbf{v}}^{(n)}). \tag{4}$$

3.3.2. Simplified-attention-based feature fusion

We remove the parameters of traditional attention mechanism [30] to propose a simplified-attention method for efficient multimodal feature fusion. We reuse $\tilde{\mathcal{V}}^{(n)}$ and $\mathcal{U}^{(n)}$ to denote the aligned feature tensor of the n_{th} image and text report. For certain w_1 and w_2, the inner product of $\mathbf{u}_{:w_1c}^{(n)}$ and $\tilde{\mathbf{v}}_{:w_2c}^{(n)}$ reflects the similarity between a text fragment and a strip image region in the c_{th} channel. Hence, multiplying the transpose of $\mathbf{U}_c^{(n)}$ with $\tilde{\mathbf{V}}_c^{(n)}$ yields a simplified attention matrix $\mathbf{A}_c^{(n)} = \left(\mathbf{U}_c^{(n)}\right)^T \cdot \tilde{\mathbf{V}}_c^{(n)}$, where element $a_{w_1w_2} = \left(\mathbf{u}_{:w_1c}^{(n)}\right)^T \cdot \tilde{\mathbf{v}}_{:w_2c}^{(n)}$

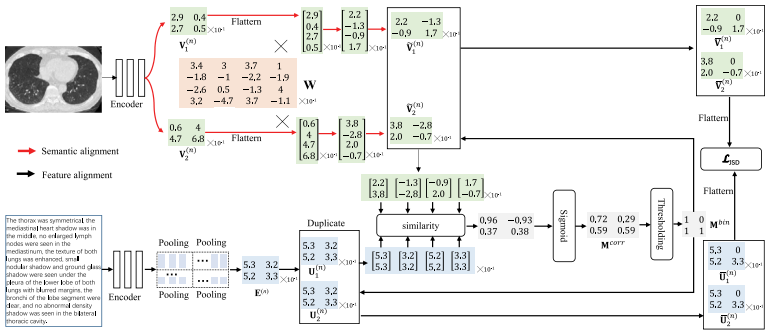

Fig. 2. An actual implementation of the IMA module, where red arrows denote semantic alignment pipeline and black arrows denote the feature alignment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

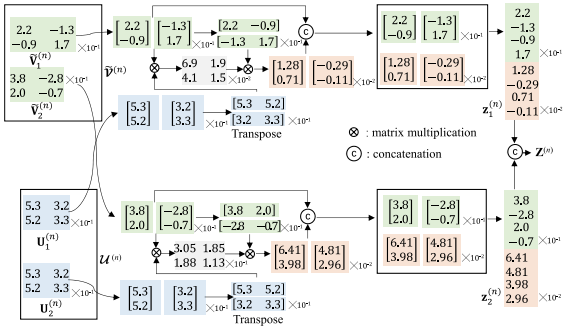

Fig. 3. An implementation process of the simplified-attention-based feature fusion.

denotes the similarity between the w_{1th} part of text and the w_{2th} image strip. Then, $\mathbf{u}^{(n)}_{:w_1c}$ can be simulated by all the strips of $\tilde{\mathbf{V}}^{(n)}_c$ using the w_{1th} row of $\mathbf{A}^{(n)}_c$ as $\tilde{\mathbf{u}}^{(n)}_{:w_1c} = \sum_{w=1}^{W} a_{w_1w} \cdot \tilde{\mathbf{v}}^{(n)}_{:wc}$. The simulation of all parts of text by image strips can be subsequently written in matrix form as $\tilde{\mathbf{U}}^{(n)}_c = \mathbf{A}^{(n)}_c \cdot \left(\tilde{\mathbf{V}}^{(n)}_c\right)^T$. We flatten $\tilde{\mathbf{V}}^{(n)}_c$ and $\tilde{\mathbf{U}}^{(n)}_c$ into vectorized form $\vec{\mathbf{v}}^{(n)}_c$ and $\vec{\mathbf{u}}^{(n)}_c$, then stack them to obtain the fused features $\mathbf{z}^{(n)}_c$ of channel c:

$$\mathbf{z}^{(n)}_c = \begin{pmatrix} \vec{\mathbf{v}}^{(n)}_c \\ \vec{\mathbf{u}}^{(n)}_c \end{pmatrix}. \tag{5}$$

Then we concatenate $\mathbf{z}^{(n)}_c$ of all channels to obtain the fused features $\mathbf{Z}^{(n)}$.

An implementation of the feature fusion is illustrated by Fig. 3, where $\tilde{\mathcal{V}}^{(n)}$ and $\mathcal{U}^{(n)}$ are provided by the IMA. We only use image features to simulate text because image contains much complex information than text and simulating image features using text features may introduce noise into image features.

3.4. Feature fusion-based classification

We use a linear classification head containing three cascaded blocks. Each block consists of a fully connected layer followed by batch normalization and Rectified Linear Unit (ReLU) activation function. The first block accepts the fused features $\mathbf{Z}^{(n)}$ as input, and the outputs of the third block are transformed to disease probabilities $\mathbf{p}^{(n)}$ through softmax function. The Cross-Entropy loss function is adopted to optimize the parameters. The prediction is depicted as:

$$\begin{aligned} \mathbf{g}^{(n)}_k &= \mathrm{ReLU}\left(\mathrm{BN}\left(\mathbf{Q}_k \cdot \mathbf{g}^{(n)}_{k-1} + \mathbf{b}_k\right)\right) \quad (k = 1, 2, 3), \\ \mathbf{p}^{(n)} &= \mathrm{softmax}\left(\mathbf{g}^{(n)}_3\right), \end{aligned} \tag{6}$$

where $\mathbf{g}^{(n)}_0 = \mathbf{Z}^{(n)}$, $\mathbf{Q}_k$ and $\mathbf{b}_k$ are weights and bias term, BN denotes batch normalization. Let $\mathbf{y}^{(n)}$ be the category label of the n_{th} instance, the Cross-Entropy loss during training can be computed as:

$$\mathcal{L}_{CE} = -\sum_{n=1}^{N}\sum_{d=1}^{D} \mathbf{y}^{(n)}(d) \log(\mathbf{p}^{(n)}(d)), \tag{7}$$

where d represents the number of diseases. The total loss function $\mathcal{L}_{total} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{JSD}$, where α and β are hyper-parameters.

4. Experiments

4.1. Datasets

The Indiana University Chest X-ray (IU X-ray) dataset[1] comprises 3955 radiology reports corresponding to 7470 anteroposterior and lateral images. Each report is made up of "Indication", "Findings", and "Impression". We use "Findings" section as textual input because it focuses on objective observations without diagnostic information. This dataset contains 119 annotated pathological categories. We construct a subset for training and testing by restricting the classification taxonomy to disease categories with 50-100 samples per class. The subset comprises 13 prevalent pathologies with 800 anteroposterior images. We do so because most categories in IU X-ray contain very few samples, which leads to trivial classification accuracy (98%) through negative prediction bias even based on single modality. We allocates $\frac{3}{8}$ of each class's samples to the training set and the remainder to testing set. For further validation, we also construct a bigger subset by restricting the taxonomy to disease categories with 10–100 samples per class. The subset comprises 43 pathologies with 1186 images. We allocates about $\frac{1}{2}$ of each class's samples to the training set and the remaining samples to the testing set. We denote this bigger subset as IU X-ray-b.

The COV-CTR dataset[2] contains 728 chest CT scans paired with structured radiology reports. Each report is made up of "Impression", "Terminology", and "Findings". Similarly, We exclusively use the "Findings" section as textual input. The COV-CTR dataset provides binary disease labels indicating the presence or absence of COVID-19 infection. Among 728 cases, 349 are COVID-19 positive, while 379 are negative samples. To establish a balanced classification benchmark, we select 200 positive and 200 negative cases as the training set. The remaining 328 cases (149 positive, 179 negative) constitute the testing set. This dataset enforces one-to-one pairing between reports and CT images.

The MIMIC-CXR-JPG dataset[3] contains 377,110 chest X-ray images associated with 227,827 medical reports. Each report may consist of "Impression", "Findings" and an equivalent section not explicitly labeled as findings or impression. We also use the "Findings" section as the textual input. Each report covers 14 diseases, and each disease

[1] https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university.
[2] https://github.com/mlii0117/COV-CTR.
[3] https://www.physionet.org/content/mimic-cxr-jpg/.

Table 1
Comparison with existing multimodal fusion methods on IU X-ray and COV-CTR Dataset.

Type	Method	IU X-ray					COV-CTR				
		Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
DF	Summation	69.91 ± 0.12	53.84 ± 0.12	16.09 ± 0.32	22.23 ± 0.45	18.66 ± 0.37	64.02 ± 0.25	65.78 ± 0.29	54.36 ± 0.51	61.83 ± 0.81	57.85 ± 0.64
	Concatenation	73.42 ± 0.14	52.83 ± 0.17	14.75 ± 0.51	17.12 ± 0.61	15.85 ± 0.56	64.33 ± 0.26	68.51 ± 0.30	53.69 ± 0.44	62.50 ± 0.51	57.76 ± 0.47
	MAF [14]	70.41 ± 0.23	53.41 ± 0.14	15.72 ± 0.65	19.45 ± 0.43	17.39 ± 0.57	66.15 ± 0.21	68.91 ± 0.33	55.03 ± 0.61	65.08 ± 0.72	59.63 ± 0.66
	LEGM [7]	74.46 ± 0.17	53.54 ± 0.22	15.64 ± 0.42	18.12 ± 0.41	16.79 ± 0.42	63.71 ± 0.28	63.56 ± 0.36	55.70 ± 0.58	61.03 ± 0.64	58.24 ± 0.61
	MATCN [18]	72.64 ± 0.21	53.43 ± 0.22	15.13 ± 0.43	20.15 ± 0.54	17.28 ± 0.48	65.85 ± 0.24	69.31 ± 0.31	55.70 ± 0.56	64.34 ± 0.64	59.71 ± 0.60
	MSAN [6]	90.01 ± 0.13	53.67 ± 0.13	7.33 ± 0.34	9.97 ± 0.42	8.45 ± 0.38	65.55 ± 0.27	65.55 ± 0.35	49.66 ± 0.56	66.07 ± 0.60	56.70 ± 0.59
	DAF [31]	74.02 ± 0.15	53.64 ± 0.12	13.51 ± 0.47	17.21 ± 0.39	15.14 ± 0.45	63.72 ± 0.22	68.56 ± 0.34	51.01 ± 0.49	62.29 ± 0.82	56.09 ± 0.63
	D3QN [19]	73.47 ± 0.27	53.11 ± 0.21	14.12 ± 0.54	16.82 ± 0.57	15.35 ± 0.56	67.07 ± 0.23	72.54 ± 0.38	54.36 ± 0.54	66.94 ± 0.64	60.00 ± 0.59
	CIM [32]	72.63 ± 0.19	53.22 ± 0.18	14.26 ± 0.38	16.79 ± 0.45	15.42 ± 0.41	65.23 ± 0.26	70.73 ± 0.32	55.03 ± 0.54	63.56 ± 0.69	58.99 ± 0.61
MF	DAM [10]	74.75 ± 0.14	53.73 ± 0.14	15.19 ± 0.36	16.22 ± 0.38	15.69 ± 0.37	63.41 ± 0.22	67.51 ± 0.35	53.69 ± 0.57	61.07 ± 0.76	57.14 ± 0.66
	MFF [9]	78.51 ± 0.24	53.13 ± 0.19	13.41 ± 0.51	21.22 ± 0.54	16.43 ± 0.55	68.90 ± 0.25	70.67 ± 0.27	51.01 ± 0.66	72.38 ± 0.70	59.84 ± 0.69
	MANN [25]	90.97 ± 0.17	50.07 ± 0.41	5.84 ± 0.46	8.76 ± 0.41	7.01 ± 0.46	66.16 ± 0.24	69.01 ± 0.37	50.33 ± 0.54	66.96 ± 0.77	57.47 ± 0.64
	MTAN [24]	72.73 ± 0.21	52.72 ± 0.35	15.92 ± 0.56	19.31 ± 0.62	17.45 ± 0.59	65.54 ± 0.25	67.54 ± 0.33	51.68 ± 0.49	65.25 ± 0.66	57.68 ± 0.56
	MSA [11]	72.35 ± 0.14	53.19 ± 0.21	15.86 ± 0.45	21.03 ± 0.48	18.08 ± 0.47	63.10 ± 0.21	66.51 ± 0.31	50.33 ± 0.60	61.48 ± 0.74	55.35 ± 0.66
	DAL [26]	91.04 ± 0.15	53.23 ± 0.23	9.56 ± 0.54	10.07 ± 0.42	9.81 ± 0.48	62.50 ± 0.24	65.82 ± 0.36	51.67 ± 0.57	60.16 ± 0.74	55.59 ± 0.65
	ReLU-Attention [33]	91.00 ± 0.11	53.14 ± 0.12	14.82 ± 0.41	23.45 ± 0.34	18.16 ± 0.41	68.29 ± 0.21	65.21 ± 0.35	54.36 ± 0.52	69.23 ± 0.76	60.90 ± 0.62
FM	MBFHA [21]	72.83 ± 0.13	53.66 ± 0.19	14.64 ± 0.52	16.22 ± 0.51	15.39 ± 0.52	64.32 ± 0.16	68.24 ± 0.30	56.37 ± 0.49	61.76 ± 0.81	58.94 ± 0.64
AM	Ours	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**

The "DF" represents direct fusion methods, "MF" denotes model-based methods, "FM" denotes feature mapping, and "AM" stands for adaptive feature mapping.

is either labeled as "Positive", "Negative", "Uncertain", or remains unlabeled (blank). In order for fast training and validation, we randomly select 1323 anteroposterior images and the corresponding reports to construct a subset, from which 600 image-text pairs are chosen for training and the rest 723 pairs are used for testing.

4.2. Experimental settings and implementation details

We compute Accuracy (Acc), Area Under the ROC Curve (AUC), Precision (P), Recall (R), and F1-score (F1) to evaluate the proposed multimodal fusion method on medical image classification task. For the binary classification on COV-CTR dataset, these metrics are computed following standard definitions. For multi-label classification on IU X-ray and IU X-ray-b datasets, we aggregate the prediction results of all labels and compute the Acc globally (similar to a single-label binary classification problem). We adopt a one-vs-rest binarization strategy to calculate the other four metrics, then macro-average across all classes. Similarly, the disease prediction on MIMIC-CXR-JPG dataset is also a multi-label classification task. We follow the officially published mention extraction task to predict each disease as mentioned ("Positive", "Negative", "Uncertain") or unmentioned (blank). Then, the identical method applied to IU X-ray dataset can be used for metric calculation. The optimizer is Adam with mini-batch size 8 and initial learning rate $\eta = 1 \times 10^{-8}$. Training terminates at 200 epochs. All experiments are conducted on an NVIDIA RTX 4090 GPU. The feature encoders are warm-started from pretrained unimodal classifiers. All the results are presented with a 95% confidence interval, reflecting statistical reliability at a 5% significance level. The best results are rendered in bold, the second best results are underlined.

4.3. Comparative experiments

Seventeen methods are used for comparison. We categorize feature summation (direct summation, MAF [14], LEGM [7]), concatenation (direct concatenation, MATCN [18], MSAN [6]), element-wise multiplication, and their combinations (DAF [31], D3QN [19], CIM [32]) as direct fusion methods. We categorize fusion by attention (DAM [10]), ReLU-attention [33], long short-term memory (MFF [9]), cross-attention (MANN [25], MTAN [24]), and transformer (MSA [11], DAL [26]) as model-based methods. MBFHA [21] exploits feature mapping (MFB pooling), and our approach can be defined as fusion based on adaptive mapping. The results on IU X-ray and COV-CTR datasets are shown in Table 1, the results on IU X-ray-b and MIMIC-CXR-JPG datasets are shown in Table 2.

We also demonstrate each method's confusion matrix of the classification results on IU X-ray-b, COV-CTR and MIMIC-CXR-JPG datasets in Fig. 4. For the binary classification on COV-CTR dataset, the confusion matrix is computed in the standard way. The true positive, true

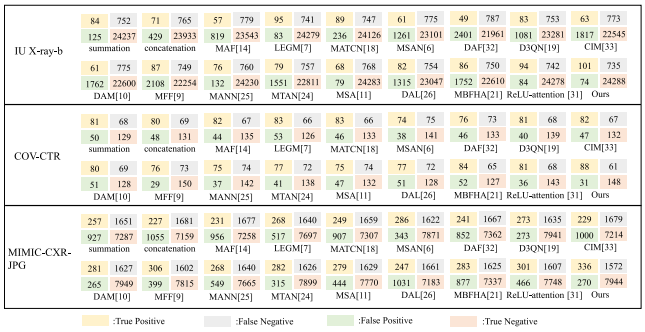

Fig. 4. The confusion matrices of our method and the comparative methods with respect to the classification results on IU X-ray-b, COV-CTR and MIMIC-CXR-JPG datasets.

negative, false positive and false negative samples add up to the total number of testing samples 328. In the multi-label classification on IU X-ray-b and MIMIC-CXR-JPG datasets, we compute the global confusion matrix using each sample's disease-level binary prediction. Therefore, the four entries of a confusion matrix add up to the number of total disease-level predictions. For IU X-ray-b, the summation is $586 \times 43 = 25{,}198$. For MIMIC-CXR-JPG, the summation is $723 \times 14 = 10{,}122$.

In addition, we record the training time of different models on IU X-ray-b, COV-CTR and MIMIC-CXR-JPG datasets in Table 3. Since the number of training sample are different for the three datasets, we record the mean time (in minute) per mini-batch required to complete 200 training epochs for each model.

4.3.1. Comparison with direct fusion

The direct fusion methods in Table 1 basically obtain close results to simply adding or concatenating the output features of image and text encoders for classification. MSAN [6] method yields obviously higher Acc than other direct fusion methods on IU X-ray dataset because MSAN method exerts extra classification loss function on feature extraction from two single modalities. The primary difference between our method and direct fusion methods is that we adopt semantic mapping to reduce the cross-modal semantic gap before feature fusion. The semantic mapping is guided by the JSD loss to guarantee cross-modal semantic consistency. Our method achieves better results on IU X-ray and COV-CTR dataset than these direct fusion methods. The lower recall and precision on IU X-ray (vs. COV-CTR) are caused by class imbalance.

The class imbalance problem is more obvious on IU X-ray-b dataset. In Table 2, the direct fusion methods have good performance because negative samples dominate the training data of each class, the models can simply predict more samples as negative to promote the Acc.

Table 2
Comparison with existing methods on IU X-ray-b and MIMIC-CXR-JPG Dataset.

Type	Method	IU X-ray-b					MIMIC-CXR-JPG				
		Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
DF	Summation	96.52 ± 0.14	52.45 ± 0.15	8.61 ± 0.63	18.60 ± 0.43	11.77 ± 0.68	74.53 ± 0.30	50.83 ± 0.35	19.99 ± 0.44	36.52 ± 0.71	25.84 ± 0.55
	Concatenation	95.26 ± 0.15	51.99 ± 0.14	7.76 ± 0.59	14.39 ± 0.52	10.08 ± 0.63	72.97 ± 0.28	50.67 ± 0.37	12.17 ± 0.45	41.03 ± 0.57	18.77 ± 0.60
	MAF [14]	93.66 ± 0.16	52.12 ± 0.17	9.04 ± 0.53	16.34 ± 0.59	11.64 ± 0.59	73.99 ± 0.35	51.26 ± 0.38	12.93 ± 0.60	54.78 ± 0.65	20.92 ± 0.84
	LEGM [7]	96.73 ± 0.11	51.67 ± 0.15	7.01 ± 0.47	18.60 ± 0.57	10.18 ± 0.58	78.69 ± 0.25	51.93 ± 0.28	7.86 ± 0.59	28.57 ± 0.61	12.33 ± 0.79
	MATCN [18]	96.10 ± 0.16	51.87 ± 0.15	7.24 ± 0.69	11.63 ± 0.66	8.92 ± 0.72	74.65 ± 0.27	50.71 ± 0.34	11.63 ± 0.50	43.12 ± 0.73	18.32 ± 0.69
	MSAN [6]	91.92 ± 0.17	51.71 ± 0.13	5.80 ± 0.56	11.86 ± 0.49	7.79 ± 0.62	80.59 ± 0.29	51.32 ± 0.33	13.21 ± 0.49	53.91 ± 0.66	21.22 ± 0.69
	DAF [31]	87.35 ± 0.16	51.36 ± 0.13	4.57 ± 0.70	7.16 ± 0.65	5.58 ± 0.72	75.11 ± 0.27	51.19 ± 0.26	18.93 ± 0.47	38.14 ± 0.78	25.30 ± 0.59
	D3QN [19]	92.72 ± 0.17	52.38 ± 0.12	7.39 ± 0.69	18.60 ± 0.62	10.58 ± 0.82	81.15 ± 0.26	51.78 ± 0.36	16.59 ± 0.68	50.18 ± 0.67	24.94 ± 0.85
	CIM [32]	89.72 ± 0.15	52.85 ± 0.14	9.95 ± 0.62	16.35 ± 0.55	12.37 ± 0.64	73.53 ± 0.31	50.74 ± 0.27	12.95 ± 0.47	50.73 ± 0.62	20.63 ± 0.65
MF	DAM [10]	89.93 ± 0.21	52.12 ± 0.11	6.05 ± 0.68	9.32 ± 0.60	7.34 ± 0.69	81.31 ± 0.24	51.87 ± 0.35	11.69 ± 0.52	39.76 ± 0.74	18.07 ± 0.70
	MFF [9]	88.66 ± 0.12	52.51 ± 0.12	7.21 ± 0.59	8.95 ± 0.65	7.99 ± 0.62	80.23 ± 0.28	51.37 ± 0.39	20.06 ± 0.59	52.67 ± 0.52	29.05 ± 0.71
	MANN [25]	96.46 ± 0.15	52.76 ± 0.12	6.71 ± 0.63	18.70 ± 0.48	9.88 ± 0.76	78.37 ± 0.24	52.31 ± 0.37	16.76 ± 0.54	42.94 ± 0.64	24.11 ± 0.66
	MTAN [24]	90.84 ± 0.14	52.89 ± 0.13	7.95 ± 0.60	18.71 ± 0.52	11.16 ± 0.69	80.82 ± 0.33	52.04 ± 0.31	12.81 ± 0.58	47.48 ± 0.88	20.18 ± 0.80
	MSA [11]	96.64 ± 0.16	51.31 ± 0.11	5.23 ± 0.65	13.95 ± 0.68	7.61 ± 0.80	79.52 ± 0.29	52.13 ± 0.40	12.72 ± 0.48	43.64 ± 0.59	19.70 ± 0.64
	DAL [26]	91.79 ± 0.15	52.34 ± 0.09	10.91 ± 0.57	9.77 ± 0.58	10.31 ± 0.58	73.40 ± 0.26	51.17 ± 0.32	14.11 ± 0.59	50.29 ± 0.79	22.04 ± 0.80
	ReLU-Attention [33]	96.72 ± 0.14	52.72 ± 0.13	9.11 ± 0.60	13.95 ± 0.46	11.02 ± 0.59	79.52 ± 0.25	52.33 ± 0.26	16.28 ± 0.44	45.78 ± 0.58	24.02 ± 0.56
FM	MBFHA [21]	90.07 ± 0.14	52.54 ± 0.10	9.21 ± 0.56	16.35 ± 0.60	11.78 ± 0.62	75.28 ± 0.27	51.97 ± 0.30	14.38 ± 0.54	28.72 ± 0.79	19.16 ± 0.66
AM	Ours	**96.79 ± 0.13**	**53.13 ± 0.15**	**11.91 ± 0.51**	**20.93 ± 0.66**	**15.18 ± 0.59**	**81.80 ± 0.27**	**52.91 ± 0.33**	**23.17 ± 0.50**	**57.14 ± 0.60**	**32.97 ± 0.61**

The "DF" represents direct fusion methods, "MF" denotes model-based methods, "FM" denotes feature mapping, and "AM" stands for adaptive feature mapping.

Table 3
Training time of different models on IU X-ray-b, COV-CTR and MIMIC-CXR-JPG Dataset.

Type	Method	Training time (minutes per batch of 200 epochs)		
		IU X-ray-b	COV-CTR	MIMIC-CXR-JPG
DF	Summation	4.4	3.8	5.3
	Concatenation	4.5	3.9	5.4
	MAF [14]	5.4	4.6	6.5
	LEGM [7]	5.2	4.4	6.2
	MATCN [18]	5.2	4.5	6.2
	MSAN [6]	5.1	4.4	6.1
	DAF [31]	5.5	4.8	6.7
	D3QN [19]	5.5	4.7	6.6
	CIM [32]	5.8	4.9	6.9
MF	DAM [10]	5.8	5.0	7.0
	MFF [9]	5.4	4.6	6.4
	MANN [25]	5.5	4.7	6.6
	MTAN [24]	5.2	4.5	6.3
	MSA [11]	5.3	4.6	6.4
	DAL [26]	5.4	4.6	6.4
	ReLU-Attention [33]	5.1	4.4	6.1
FM	MBFHA [21]	5.9	5.1	7.1
AM	Ours	4.8	4.2	5.8

The "DF" represents direct fusion methods, "MF" denotes model-based methods, "FM" denotes feature mapping, and "AM" stands for adaptive feature mapping.

This can be observed in the confusion matrices in Fig. 4. In spite of this, our method outperforms all the direct fusion methods. The samples per class are more balanced in MIMIC-CXR-JPG dataset, the models should be well designed for good performance. Therefore, direct fusion methods are better than simple summation or concatenation. D3QN [19] achieves good results on COV-CTR and MIMIC-CXR-JPG datasets, indicating that combining multiple direct fusion methods can be effective. Our method also outperforms all the direct fusion methods on MIMIC-CXR-JPG dataset.

Table 3 indicates that though our method needs to estimate the mapping matrix, the training is still faster than the direct fusion methods that adopt complex strategy to combine the basic adding, concatenation and multiplication. The feedforward computations also consume considerable time.

4.3.2. Comparison with model-based fusion

The second block of Table 1 reveals that our method achieves better results than these model-based methods on IU X-ray and COV-CTR datasets. The model-based methods try to align all the features of each modality, but some features may have very weak cross-modality correlation. In comparison, we adopt binary mask learned with JSD loss to query strongly correlated features to perform cross-modal alignment. The MANN [25] and DAL [26] methods achieve good results on IU X-ray dataset partly because they adopt bi-directional cross-attention to enhance multimodal fusion. ReLU-Attention method [33] is inferior to our approach because it ignores the negative products of querying and key features in computing the attention weights, which could actually be meaningful.

The second block of Table 2 shows the results of model-based methods on IU X-ray-b and MIMIC-CXR-JPG datasets. The model-based methods cannot outperform the direct fusion methods on IU X-ray-b dataset because complex models are susceptible to overfitting on imbalanced dataset. The model-based approaches perform better on MIMIC-CXR-JPG dataset. This indicates that complex multimodal fusion methods benefit classification on balanced dataset. Our method still outperforms the model-based methods on the two datasets.

The training time of the model-based methods is shown in Table 3. These models generally have large number of parameters. In contrast, our method only needs to estimate the semantic mapping matrix. Since we define the size of feature map as 7×7, the number of parameters to be estimated is $49 \times 49 = 2,401$, which is much less than cross-attention and a standard transformer. Accordingly, the training speed of our method is faster than the model-based methods.

4.3.3. Comparison with feature mapping-based fusion

The results of mapping-based fusion method are shown in the third block of Tables 1 and 2. The MBFHA [21] method learns a projection

Table 4
Ablation study about multimodal alignment on IU X-ray and COV-CTR Dataset.

I	R	RM	IM	JSD	Msk	Bin	IU X-ray					COV-CTR				
							Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
✓							86.65 ± 0.23	51.13 ± 0.22	10.16 ± 0.45	15.44 ± 0.31	12.26 ± 0.43	67.07 ± 0.26	71.47 ± 0.31	55.70 ± 0.58	66.40 ± 0.67	60.58 ± 0.62
	✓						67.87 ± 0.21	53.16 ± 0.21	15.14 ± 0.42	17.59 ± 0.41	16.27 ± 0.42	61.28 ± 0.25	64.35 ± 0.35	54.36 ± 0.65	57.86 ± 0.69	56.06 ± 0.67
✓	✓						83.69 ± 0.13	52.01 ± 0.23	7.31 ± 0.34	8.92 ± 0.45	8.04 ± 0.39	66.16 ± 0.20	69.73 ± 0.33	53.02 ± 0.62	65.83 ± 0.84	58.73 ± 0.71
✓	✓			✓			78.49 ± 0.12	50.79 ± 0.34	6.61 ± 0.51	8.93 ± 0.56	7.60 ± 0.54	66.77 ± 0.25	70.57 ± 0.32	52.34 ± 0.71	67.24 ± 0.61	58.86 ± 0.68
✓	✓			✓	✓		83.67 ± 0.15	51.67 ± 0.31	9.09 ± 0.41	9.11 ± 0.31	9.10 ± 0.36	67.07 ± 0.24	69.72 ± 0.33	54.36 ± 0.66	66.94 ± 0.76	60.00 ± 0.71
✓	✓			✓	✓	✓	78.52 ± 0.22	51.06 ± 0.13	6.11 ± 0.54	9.91 ± 0.66	7.56 ± 0.61	67.38 ± 0.23	70.41 ± 0.32	50.33 ± 0.52	69.44 ± 0.81	58.36 ± 0.64
✓	✓	✓					81.79 ± 0.13	51.76 ± 0.14	13.63 ± 0.66	7.71 ± 0.56	9.85 ± 0.63	66.16 ± 0.27	73.41 ± 0.42	55.03 ± 0.56	65.07 ± 0.97	59.63 ± 0.74
✓	✓	✓		✓			81.09 ± 0.32	53.41 ± 0.21	15.69 ± 0.41	8.90 ± 0.47	11.36 ± 0.49	66.46 ± 0.24	73.32 ± 0.28	46.98 ± 0.69	69.31 ± 0.88	56.00 ± 0.78
✓	✓	✓		✓	✓		88.01 ± 0.19	51.43 ± 0.16	12.46 ± 0.56	7.78 ± 0.45	9.58 ± 0.51	67.38 ± 0.26	70.92 ± 0.36	52.35 ± 0.82	68.42 ± 0.86	59.32 ± 0.85
✓	✓	✓		✓	✓	✓	84.54 ± 0.13	53.49 ± 0.16	13.30 ± 0.41	16.81 ± 0.45	14.85 ± 0.43	68.29 ± 0.26	74.36 ± 0.35	58.39 ± 0.67	67.44 ± 0.85	62.59 ± 0.75
✓	✓		✓				89.15 ± 0.24	51.71 ± 0.24	11.51 ± 0.61	9.87 ± 0.62	10.63 ± 0.62	70.12 ± 0.24	72.84 ± 0.26	54.36 ± 0.60	72.97 ± 0.62	62.31 ± 0.62
✓	✓		✓	✓			90.98 ± 0.11	52.63 ± 0.15	13.64 ± 0.56	11.32 ± 0.57	12.37 ± 0.57	68.29 ± 0.22	71.89 ± 0.35	53.69 ± 0.73	69.56 ± 0.71	60.60 ± 0.73
✓	✓		✓	✓	✓		90.95 ± 0.13	51.42 ± 0.23	7.25 ± 0.45	10.20 ± 0.65	8.48 ± 0.53	69.21 ± 0.28	73.34 ± 0.29	55.70 ± 0.63	70.34 ± 0.93	62.17 ± 0.76
✓	✓		✓	✓	✓	✓	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**

The "I" and "R" represents image and text report respectively, "IM" and "RM" means applying semantic mapping **W** to images or text reports. "JSD" denotes using JSD loss, "Msk" means applying mask to features, and "Bin" represents thresholding the mask to binary digits.

for each modality and computes the outer product of the projected features of both modalities to achieve multimodal fusion. MBFHA method also employs all the features to perform cross-modal correlation and is susceptible to the negative influence of weakly correlated features. In addition, the bilinear mapping method does not employ the distribution alignment scheme, this will further degrade its performance. In consequence, the Acc of our method exceeds that of MBFHA method by over 18 percents and 7 percents on IU X-ray and COV-CTR dataset respectively, and the advantage of our method on either IU X-ray-b or MIMIC-CXR-JPG datasets surpasses 6 percents. The training time of MBFHA is recorded in Table 3. It is surprisingly long because MBFHA adopts hybrid attention-based reconstruction to promote the feature quality. The training time of our method is shorter for 1.1, 0.9 and 1.3 min per batch than that of MBFHA on three datasets. This validates the training efficiency of our method.

4.4. Ablation experiments about multimodal alignment

This section evaluates the classification results of our approach under different configurations of multimodal alignment methods. We provide the results of unimodal classification (using only image or text report) as baseline. We use the simplified-attention method proposed in this paper to fuse all the image and text features (with optional semantic mapping), and the results on IU X-ray and COV-CTR datasets are shown in Table 4. Since our multimodal alignment module consists of semantic alignment and feature alignment, we divide this ablation study into three groups, i.e. no semantic mapping (alignment), applying semantic mapping to text report (text semantic mapping), and applying semantic mapping to image (image semantic mapping). Within each group, we further explore the results of incrementally superposing JSD loss, correlation mask and binary mask.

4.4.1. Multimodal alignment without semantic mapping

The results are shown in the second block of Table 4. For IU X-ray dataset, without semantic mapping, The Acc stands between 78.49±0.12% and 83.69±0.13%, which is even inferior to image-based classification. Under large semantic gap, the image and text features interfere with each other, which reduces the quality of feature fusion. Exploiting only JSD loss is not effective because forcibly aligning two modalities with large semantic discrepancy will degrade the quality of feature fusion. This situation can be improved by applying JSD loss to features weighted by the correlation mask. The mask assigns different weights to features in multimodal alignment to reduce the influence of irrelevant features. Thresholding the mask to binary mask decreases the Acc because binary mask retrieves less features for multimodal alignment, which are insufficient to mitigate the semantic gap when semantic mapping is missing. For COV-CTR dataset, the Acc is more stable because the performance of two unimodal classifiers is closer to each other. The recall and precision are much higher than those on IU X-ray dataset because COV-CTR dataset has abundant and balanced training samples per class.

4.4.2. Multimodal alignment with only text semantic mapping

The results are shown in the third block of Table 4. For IU X-ray dataset, applying semantic mapping to text reports yields Acc of 81.79±0.13%, lower than directly fusing image and text features. This operation may introduce external noise into text features. Appending JSD loss to this semantic mapping further lowers Acc to 81.09±0.32%, this can be explained by the forcible alignment between image features and noisy text features. When we weight the features through correlation mask and apply JSD loss to the features, the Acc rises to 88.01±0.19%. The mask helps to reduce the noise to improve the feature fusion. However, replacing the correlation mask with binary mask decreases the Acc from 88.01±0.19% to 84.54±0.13%. The selected features from noisy text representations do not correspond to image features well. The experimental results on COV-CTR dataset exhibit similar trend to the results in Section 4.4.1.

4.4.3. Multimodal alignment with only image semantic mapping

The results are shown in the fourth block of Table 4. For IU X-ray dataset, applying semantic mapping to images obtains Acc of 89.15±0.24%. This mapping achieves satisfactory semantic alignment while preserves key image information to classification. Superposing JSD loss or mask with JSD loss on this mapping further improves the Acc. The JSD loss guides the semantically aligned image and text features to approach each other. The complete method with image-to-text mapping and binary mask with JSD loss yields the highest Acc of 91.14±0.19%. In addition, the complete method achieves the best AUC, recall, precision, and F1-score. For COV-CTR dataset, the complete method also achieves the best results. These experiments justify the efficacy of our multimodal alignment module.

4.5. Ablation experiments about feature fusion

These experiments aim to determine the most appropriate feature representations that can be fused to maximize the classification performance. We use different combinations of operations proposed by our method to generate various features for image and text, and evaluate the fused features by the simplified-attention method through classification. Specifically, we separate these experiments into two groups. In the first group, we perform multimodal alignment with JSD loss. In feature fusion stage, we combine different features of both modalities, which include all the text features, masked text features, binary-masked text features, all the image features (after semantic mapping), masked image features (after semantic mapping), and binary-masked image features (after semantic mapping). In the second group, we remove JSD loss, which means the semantic mapping and mask learning is only guided by the classification loss. Then, in feature fusion, we adopt similar combination strategy to group one.

Table 5
Comparison of our method and other fusion methods with JSD Loss.

Features to be fused	JSD	IU X-ray					COV-CTR				
		Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
Image: A+M; Text: all	w/	77.73 ± 0.21	51.62 ± 0.23	12.84 ± 0.35	11.93 ± 0.47	12.37 ± 0.42	65.24 ± 0.37	61.21 ± 0.35	38.25 ± 0.82	72.15 ± 0.86	50.00 ± 0.91
Image: A+M; Text: M	w/	77.01 ± 0.15	53.03 ± 0.21	14.01 ± 0.22	16.81 ± 0.45	15.28 ± 0.32	65.85 ± 0.34	69.12 ± 0.31	40.26 ± 0.71	72.29 ± 0.84	51.72 ± 0.80
Image: A+BM; Text: all	w/	68.02 ± 0.14	53.73 ± 0.31	14.63 ± 0.56	23.05 ± 0.42	17.90 ± 0.55	65.85 ± 0.35	73.92 ± 0.29	44.97 ± 0.80	70.53 ± 0.89	54.92 ± 0.87
Image: A+BM; Text: BM	w/	67.12 ± 0.21	53.14 ± 0.15	15.78 ± 0.45	21.17 ± 0.56	18.08 ± 0.50	66.15 ± 0.30	73.57 ± 0.37	40.26 ± 0.76	73.17 ± 0.78	51.94 ± 0.83
Image: A; Text: M	w/	91.07 ± 0.12	50.73 ± 0.17	8.51 ± 0.54	9.16 ± 0.40	8.82 ± 0.48	68.90 ± 0.31	73.83 ± 0.37	49.66 ± 0.60	73.26 ± 0.79	59.19 ± 0.68
Image: A; Text: BM	w/	90.98 ± 0.13	50.37 ± 0.11	8.76 ± 0.61	7.89 ± 0.51	8.30 ± 0.56	69.21 ± 0.37	70.75 ± 0.36	54.36 ± 0.73	71.05 ± 0.85	61.59 ± 0.79
Image: A; Text: all (ours)	w/	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**

The "A" means semantic aligning by W, "M" means masking, "A+M" means aligning with masking, "BM" means binary-masking, "A+BM" means aligning with binary-masking.

Table 6
Comparison of Our method and other fusion methods without JSD loss.

Features to be fused	JSD	IUX-ray					COV-CTR				
		Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
Image: A+M; Text: all	w/o	77.52 ± 0.13	53.35 ± 0.23	11.76 ± 0.45	10.93 ± 0.55	11.33 ± 0.50	67.37 ± 0.28	68.97 ± 0.27	43.62 ± 0.93	73.89 ± 0.71	54.86 ± 0.93
Image: A+M; Text: M	w/o	77.86 ± 0.21	53.17 ± 0.23	10.12 ± 0.51	15.82 ± 0.61	12.34 ± 0.57	68.90 ± 0.33	64.51 ± 0.31	51.68 ± 0.76	71.96 ± 0.81	60.16 ± 0.80
Image: A+BM; Text: all	w/o	68.34 ± 0.14	53.57 ± 0.25	16.18 ± 0.41	21.72 ± 0.46	18.55 ± 0.44	67.37 ± 0.35	71.47 ± 0.33	47.65 ± 0.77	71.00 ± 0.78	57.03 ± 0.81
Image: A+BM; Text: BM	w/o	68.07 ± 0.23	53.64 ± 0.21	15.81 ± 0.53	20.97 ± 0.51	18.03 ± 0.53	66.77 ± 0.34	71.63 ± 0.31	46.31 ± 0.82	70.41 ± 0.78	55.87 ± 0.84
Image: A; Text: M	w/o	90.02 ± 0.14	50.31 ± 0.21	8.29 ± 0.46	9.24 ± 0.52	8.74 ± 0.49	69.51 ± 0.38	69.51 ± 0.38	53.02 ± 0.84	72.47 ± 0.95	61.24 ± 0.90
Image: A; Text: BM	w/o	90.96 ± 0.16	50.13 ± 0.28	9.72 ± 0.57	12.91 ± 0.58	11.09 ± 0.59	67.07 ± 0.36	71.73 ± 0.35	46.97 ± 0.76	70.71 ± 0.83	56.45 ± 0.81
Image: A; Text: all	w/o	89.15 ± 0.24	51.71 ± 0.24	11.51 ± 0.61	9.87 ± 0.62	10.63 ± 0.62	70.12 ± 0.24	72.84 ± 0.26	54.36 ± 0.60	72.97 ± 0.62	62.31 ± 0.62
Image: A; Text: all (ours)	w/	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**

The "A" means semantic aligning by W, "M" means masking, "A+M" means aligning with masking, "BM" means binary-masking, "A+BM" means aligning with binary-masking.

4.5.1. Comparison of our method and other fusion methods with JSD loss

The classification results through feature fusion with JSD loss are shown in Table 5. For IU X-ray dataset, when correlation mask is applied to image features, the Acc is lower than 78%, regardless of whether the text features are masked. When we further reduce the image features through binary mask, the Acc continues to drop to less than 69%. Applying mask to text features individually does not encounter this problem. The images of this dataset contain more information than text reports, some information is irrelevant to text features but still beneficial to the classification, so removing these features degrades the classification performance. For COV-CTR dataset, the Acc shows less fluctuation due to the balanced information of images and texts. Our feature fusion module adopts all the aligned features for classification, and yields the best results on the two datasets. Using more image features for multimodal fusion benefits the subsequent classification.

4.5.2. Comparison of our method and other fusion methods without JSD loss

The classification results through feature fusion without JSD loss are shown in Table 6. We also append the results of our method in this table for comparison. The experimental results exhibit very similar trend to that in Table 5. For IU X-ray dataset, masking image features leads to significant decrease of performance while masking text features individually does not. The results on COV-CTR dataset shows patterns consistent with IU X-ray dataset. The results of all the feature fusion methods without JSD loss are inferior to the results of our method. This also justifies the importance of image features to subsequent classification.

4.6. Ablation experiments about hyper-parameters

4.6.1. Sensitivity to α and β in loss function $\mathcal{L}_{total}$

In the loss function $\mathcal{L}_{total} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{JSD}$, The hyper-parameters α and β balance the contribution of $\mathcal{L}_{CE}$ and $\mathcal{L}_{JSD}$ to $\mathcal{L}_{total}$. Since the sum of α and β is 1, we only vary the values of α and evaluate the influence to classification results. We set α as 0.10, 0.30, 0.50, 0.70, 0.99, and 1 (without JSD loss) respectively. The results on IU X-ray and COV-CTR datasets are shown in Table 7.

With the increase of α, the Acc rises gradually and reaches the peaks when α is set to 0.99. When α is equal to 1, i.e. only classification loss is adopted, the Acc shows observable reduction. This validates the efficacy of the JSD loss function in enhancing model performance. Owing to the heterogeneity of image and text, the image and text features have significant semantic divergence, which results in very large JSD loss value. To balance the JSD loss, β should be very small. This regularity is applicable to diverse data. This explains why $\beta = 0.01$ results in the best results. In our experiments, we set $\alpha = 0.99$.

4.6.2. Sensitivity to the thresholds for deriving $\mathbf{M}^{bin}$

This section tests the influence of different thresholds for deriving $\mathbf{M}^{bin}$ to the classification. We vary the threshold from 0.1 to 0.9, the classification results are shown in Table 8. For IU X-ray dataset, our method achieves the best results at threshold 0.5. When the threshold is set as 0.1, the performance shows noticeable decrease. For COV-CTR dataset, we achieves the highest performance at threshold 0.3. The optimal threshold for IU X-ray dataset is larger than that for COV-CTR dataset. The image and text information of IU X-ray data is more imbalanced, so we should select features with stronger inter-modality correlations to enhance the multimodal alignment. For data with balanced information (COV-CTR), a smaller threshold can achieve satisfactory performance. Hence, we set the threshold as 0.5 for IU X-ray dataset, and 0.3 for COV-CTR dataset.

4.6.3. Sensitivity to the size of feature map $\mathbf{V}_c^{(n)}$

This section tests the influence of different feature sizes to the multimodal classification performance. In our method, we pool the text feature matrix to fit the size of image feature map, so we only evaluate different sizes of the image feature map that will be fed to multimodal fusion module. We vary the size of feature map as 2×2, 3×3, 5×5, 7×7, and 9×9, and conduct multimodal classification, the results are shown in Table 9. The default feature map size 7×7 leads to the highest classification performance. When feature map is smaller than 5×5, the Acc on IU X-ray dataset drops significantly. The original image in IU X-ray dataset has 2048×2048 resolution, therefore 2×2 or 3×3 sized feature map may fail to capture sufficient features. The images in COV-CTR dataset are much smaller, therefore the performance difference between large and small feature map is also small. We think the default feature size 7×7 of VGG-11 can be applied to various sized images. Hence, we set the size of feature map as 7×7.

4.7. Visualization

4.7.1. Visualization of our model with different configurations

We follow Grad-CAM method [34] to compute a heatmap for each medical image using the gradient produced by diagnosis label in classification, and superpose the heatmap onto the image. The heatmap shows the focus of attention of the model. Fig. 5 shows one of our visualization cases of COV-CTR dataset, where the top row shows the original image and its radiology report, the middle row shows the heatmaps generated by different configurations, and the bottom row superposes the heatmaps to the original image.

The left panel of the top row displays the original image with the lesion areas highlighted by red arrows. In the right panel, different

Table 7
Experimental results with different values of α in $\mathcal{L}_{total}$.

α	IU X-ray					COV-CTR				
	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
0.10	76.69 ± 0.23	52.60 ± 0.15	9.81 ± 0.54	20.97 ± 0.65	13.37 ± 0.64	67.68 ± 0.35	71.45 ± 0.37	46.31 ± 0.54	72.63 ± 0.64	56.56 ± 0.60
0.30	88.52 ± 0.25	53.25 ± 0.16	10.26 ± 0.48	23.32 ± 0.56	14.25 ± 0.57	68.90 ± 0.24	74.08 ± 0.34	48.99 ± 0.52	73.74 ± 0.87	58.87 ± 0.65
0.50	89.45 ± 0.19	51.94 ± 0.22	6.03 ± 0.53	10.20 ± 0.67	7.58 ± 0.60	69.51 ± 0.29	71.04 ± 0.36	51.01 ± 0.43	73.79 ± 0.83	60.32 ± 0.58
0.70	90.61 ± 0.21	51.98 ± 0.16	7.87 ± 0.68	16.96 ± 0.66	10.75 ± 0.77	69.21 ± 0.23	73.92 ± 0.28	50.33 ± 0.51	73.52 ± 0.81	59.75 ± 0.63
0.99	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**
1.00	89.15 ± 0.24	51.71 ± 0.14	11.51 ± 0.61	9.87 ± 0.62	10.63 ± 0.62	70.12 ± 0.24	72.84 ± 0.26	54.36 ± 0.60	72.97 ± 0.62	62.31 ± 0.62

Table 8
Experimental results with different thresholds of deriving $\mathbf{M}^{bin}$.

Threshold	IU X-ray					COV-CTR				
	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
0.1	89.89 ± 0.23	50.08 ± 0.17	8.31 ± 0.43	19.23 ± 0.56	11.61 ± 0.52	68.29 ± 0.29	70.91 ± 0.33	49.66 ± 0.51	71.84 ± 0.66	58.73 ± 0.58
0.3	91.10 ± 0.19	51.89 ± 0.21	11.46 ± 0.57	15.36 ± 0.59	13.13 ± 0.59	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**
0.5	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	68.29 ± 0.28	73.25 ± 0.35	46.98 ± 0.63	73.68 ± 0.78	57.38 ± 0.71
0.7	90.95 ± 0.17	50.91 ± 0.23	10.14 ± 0.59	13.15 ± 0.68	11.45 ± 0.63	67.99 ± 0.31	71.45 ± 0.28	56.37 ± 0.46	67.74 ± 0.74	61.53 ± 0.58
0.9	90.79 ± 0.20	50.79 ± 0.19	7.41 ± 0.61	17.56 ± 0.66	10.42 ± 0.72	64.63 ± 0.37	72.24 ± 0.31	53.69 ± 0.52	62.50 ± 0.72	57.76 ± 0.61

Table 9
Experimental results with different sizes of feature map $\mathbf{V}_c^{(n)}$.

$H \times W$	IU X-ray					COV-CTR				
	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
2 × 2	74.12 ± 0.22	53.32 ± 0.18	11.51 ± 0.36	17.36 ± 0.47	13.84 ± 0.41	63.11 ± 0.29	64.35 ± 0.36	35.57 ± 0.72	67.95 ± 0.86	46.70 ± 0.82
3 × 3	77.71 ± 0.12	51.23 ± 0.21	7.90 ± 0.48	18.50 ± 0.48	11.07 ± 0.56	63.41 ± 0.24	65.23 ± 0.28	48.32 ± 0.74	62.61 ± 0.73	54.54 ± 0.75
5 × 5	90.98 ± 0.20	51.19 ± 0.23	7.69 ± 0.52	15.08 ± 0.57	10.19 ± 0.59	65.24 ± 0.25	66.04 ± 0.29	53.02 ± 0.67	64.22 ± 0.86	58.09 ± 0.75
7 × 7	**91.14 ± 0.19**	**53.95 ± 0.12**	**16.43 ± 0.48**	**25.41 ± 0.53**	**19.96 ± 0.52**	**71.95 ± 0.21**	**74.93 ± 0.34**	**59.06 ± 0.59**	**73.95 ± 0.68**	**65.67 ± 0.63**
9 × 9	88.69 ± 0.21	51.56 ± 0.12	7.75 ± 0.45	16.04 ± 0.57	10.45 ± 0.53	63.11 ± 0.40	64.20 ± 0.32	55.03 ± 0.60	60.29 ± 0.64	57.54 ± 0.62

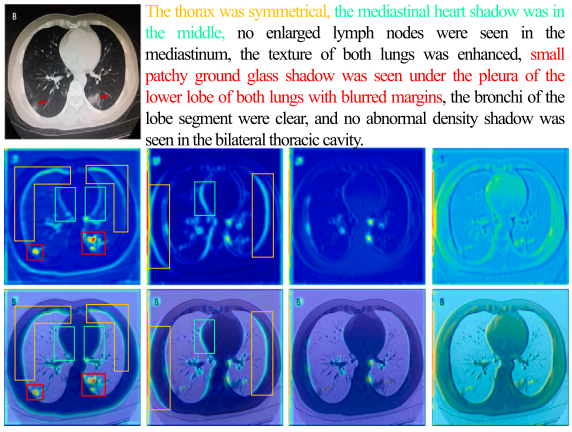

Fig. 5. The visualization results, (a) our approach with semantic mapping **W**, binary mask $\mathbf{M}^{bin}$, and JSD loss, (b) removing **W** from our model, (c) removing the mask $\mathbf{M}^{bin}$ and JSD loss from our model, (d) removing the mask $\mathbf{M}^{bin}$ from our model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sections of the text report are color-coded to correspond with specific anatomical regions. The thorax description is marked in orange, the mediastinal heart shadow description is marked in light blue, and the ground-glass shadow description is marked in red. This facilitates a clear reference between the image and the textual analysis. The middle row shows four heatmaps generated by our model with different network configurations. Areas of high attention should be indicated by warmer colors. This helps to understand which parts of the image the network focuses on. Manual annotations in the form of lines and rectangular boxes are overlaid on the heatmap, color-matched to the text descriptions in the report. This helps to correlate the network's focus areas with the regions described in the report.

The four heatmaps in the middle row are captioned with (a), (b), (c) and (d). Heatmap (a) is generated by our approach. We only remove **W** from our model to generate heatmap (b). We remove the mask and JSD loss to generate heatmap (c). In this case, all the text features and semantically aligned image features are sent to the simplified-attention module for feature fusion. We only remove the mask and preserve **W** and JSD loss to generate heatmap (d). In the last case, the JSD loss is applied to all text features and semantically aligned image features, which are fed to the simplified-attention module for feature fusion.

It is clear that our approach achieves the best correspondence between image and text (Fig. 5(a)). The thorax (in the orange wire-frames), mediastinal region around heart (in the light blue rectangles), and the lesions (in the red rectangles) are accurately highlighted in the heatmap by our model. These observations are also depicted by radiology report. The lesions strongly support the diagnosis of COVID-19 infection. Removing **W** from our approach (Fig. 5(b)) causes the overlook of the lesions. The semantic alignment ability is missing without **W**. But the model still notices the thorax and mediastinal region owing to the feature alignment by mask and JSD loss. Removing mask from our approach (Fig. 5(c) and (d)) leads to worse results because the semantic alignment alone is too weak to correlate image and text well. Fig. 5 indicates that our approach corresponds image and text well and is sensitive to key information to diagnosis.

4.7.2. Visualized comparison of our method with other methods

We demonstrate the visualization results of our method and several competitive methods on two medical images in Fig. 6, where the left and right panel corresponds to one of the two examples respectively. Each panel shows the original report, the original image, the heatmaps of our method and MSAN [6], MFF [9], DAL [26], D3QN [19], LEGM [7] methods. The description about the pathological region is marked in red in the report, and the lesion area in the original image has been annotated with red arrows that denote the ground truth. It is clear that our model can focus on the lesions while the competitive methods perform worse. Specifically, MSAN, DAL and D3QN methods fail to find any suspicious lesion area, MFF methods erroneously predicts the mediastinal region as a lesion area, LEGM method actually focuses on the red arrows but not the lesion area. We deem the advantage of our method lies in the semantic mapping and feature alignment guided by JSD loss in our model.

4.8. Limitations of our method

4.8.1. Dependency on dual modalities

Our method needs the image and text to be simultaneously available. If one of them is missing, the classification results will be poor.

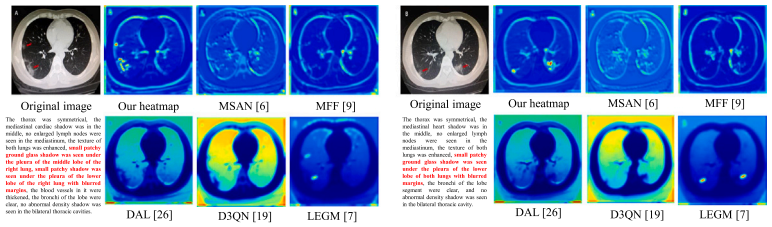

Fig. 6. Visualized comparison of our method with other methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Experimental results with missing modality on IU X-ray and COV-CTR dataset.

Modality		IU X-ray					COV-CTR				
Image	Text	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)	Acc (%)	AUC (%)	R (%)	P (%)	F1 (%)
w/o	w/	63.61 ± 0.12	49.97 ± 0.13	6.09 ± 0.32	5.23 ± 0.43	5.63 ± 0.39	59.75 ± 0.30	64.22 ± 0.31	54.36 ± 0.49	55.86 ± 0.89	55.10 ± 0.69
w	w/o	60.97 ± 0.13	50.03 ± 0.10	5.71 ± 0.41	5.13 ± 0.46	5.40 ± 0.44	56.40 ± 0.39	64.52 ± 0.29	45.63 ± 0.51	52.30 ± 0.97	48.74 ± 0.71

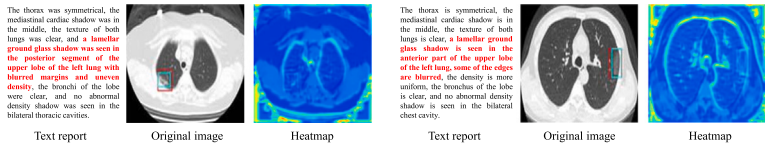

Fig. 7. Examples of unsatisfactory multimodal alignment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To show this limitation, we use only image or text modality to train our model and conduct classification. To enable multimodal feature fusion, we replace the logically missing modality with zeros in the experiments. Owing to the bias terms, the feature extraction results from zero input will not be zero. The experimental results on IU X-ray and COV-CTR datasets can be seen in Table 10, where the available modality is denoted by "w/" while the unavailable modality is denoted by "w/o". The method fails to achieve good classification performance when one modality is missing.

4.8.2. Inability to achieve fine-grained image-text alignment

In our multimodal alignment module, the binary mask is learned from holistic image feature maps and text feature matrices, and applied to these holistic features. This strategy assumes that the feature at certain location of the image map naturally corresponds to the feature at the same location of the text feature matrix, which may not be true. Fine-grained image-text matching is necessary [35]. Fig. 7 exhibits two examples on which our method does not yield satisfactory multimodal alignment. The left and right panel corresponds to one of the two examples respectively, where the original report, original image and the computed heatmap are shown. The description about the pathological region is also marked in red in the report, and the lesion area in the original image is annotated with colored rectangles that denote the ground truth. It is clear that the model does not focus on the lesions. After feature aggregation, the lesion features locate at the edge of the feature map. But the text features of pathological region approximately locate in the middle of the report. The location mismatch problem causes the unsatisfactory alignment. We will address the issue in future work.

5. Conclusion

We propose an image-text fusion model that consists of multimodal alignment and feature fusion modules. The multimodal semantic alignment is realized by a linear mapping from image to text and multimodal feature alignment is achieved through binary correlation mask with JSD loss. We use parameter-free simplified-attention to conduct feature fusion. The method promotes the feature fusion quality to improve the subsequent multimodal classification. This method's strength lies in three aspects. First, the semantic mapping resolves modality heterogeneity by aligning medical image to text features while preserving diagnostically critical image information. Second, the binary mask with JSD loss selects disease-relevant multimodal features to optimize the feature alignment and provide clinical interpretability. At last, the parameter-free simplified-attention mechanism is computationally efficient while maintaining accuracy in small medical datasets.

This work still has two limitations. First, our method needs both image and text modalities to be available. If one of the two modalities is missing, the subsequent classification cannot be satisfactory. Second, the binary mask is learned using the holistic features of image and text. It assumes the image feature at certain location of the feature map corresponds well to the text feature at the same location in the text feature matrix, but this might not be true. The limitations have been discussed in detail with examples in Section 4.8.

We plan to solve the two problems in the future. By replacing the VGG network by vision transformers, the image can also be represented by tokens. Then, we can perform token-level image-text feature alignment. Specifically, we compute the similarity between a text token and all the image tokens, then use the image tokens with large similarity to simulate the text token. By excluding dissimilar image tokens, the irrelevant cross-modal information can be removed. After the fine-grained multimodal alignment, text and image features can be used for classification individually. In this way, even one of the modalities is unavailable, the classification can also be conducted. We plan to apply the improved approach to address more medical tasks, e.g. detecting breast lesion based on mammography images and radiology reports, to facilitate the clinical diagnosis.

CRediT authorship contribution statement

Jian Zhang: Writing – original draft, Supervision, Conceptualization. **Kaihao He:** Software, Data curation. **Zunlei Feng:** Funding acquisition, Formal analysis. **Shuifa Sun:** Supervision, Methodology, Investigation. **Xiaoyan Sun:** Writing – review & editing, Software. **Zhenming Yuan:** Writing – review & editing, Visualization. **Jun Yu:** Writing – review & editing, Supervision.

Code availability

https://github.com/busitl/Self-Adaptive-Image-Text-Fusion

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

This work is supported by Natural Science Foundation of China, No. 62376248, and by Information Technology Center, ZheJiang University.

Data availability

I have shared the link to my code in the manuscript.

References

[1] X. He, Y. Wang, S. Zhao, X. Chen, Co-attention fusion network for multimodal skin cancer diagnosis, Pattern Recognit. 133 (2023) 108990.
[2] S. Deng, X. Zhang, S. Jiang, A diagnostic report supervised deep learning model training strategy for diagnosis of COVID-19, Pattern Recognit. 149 (2024) 110232.
[3] F. Zhao, C. Zhang, B. Geng, Deep multimodal data fusion, ACM Comput. Surv. 56 (9) (2024) 216.
[4] R. Hou, D. Zhou, R. Nie, D. Liu, X. Ruan, Brain CT and MRI medical image fusion using convolutional neural networks and a dual-channel spiking cortical model, Med. Biol. Eng. Comput. 57 (2019) 887–900.
[5] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, H. Roth, Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation, Med. Image Anal. 65 (2020) 101766.
[6] X. He, Y. Deng, L. Fang, Q. Peng, Multi-modal retinal image classification with modality-specific attention network, IEEE Trans. Med. Imaging 40 (6) (2021) 1591–1602.
[7] W. Li, Y. Zhang, G. Wang, Y. Huang, R. Li, DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion, Biomed. Signal. Process. 80 (2023) 104402.
[8] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1839–1848.
[9] J. Li, L. Huang, C. Zhu, S. Zhang, Q. Li, Multimodal feature hierarchical fusion for text-image person re-identification, in: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, 2024, pp. 468–481.
[10] M. Zhang, X. Sun, J. Liu, S. Xu, Y. Piao, H. Lu, Asymmetric two-stream architecture for accurate RGB-D saliency detection, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 374–390.
[11] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, Automated radiographic report generation purely on transformer: A multicriteria supervised approach, IEEE Trans. Med. Imaging 41 (10) (2022) 2803–2813.
[12] X. Hou, Y. Bai, Y. Xie, Y. Zhang, L. Fu, Y. Li, C. Shang, Q. Shen, Self-supervised multimodal change detection based on difference contrast learning for remote sensing imagery, Pattern Recognit. 159 (2025) 111148.
[13] H. Liu, C. Li, Q. Wu, Y. Lee, Visual instruction tuning, in: Proceedings of the International Conference on Neural Information Processing Systems, 2023, pp. 34892–34916.
[14] Y. Chang, Z. Zheng, Y. Sun, M. Zhao, Y. Lu, Y. Zhang, DPAFNet: A residual dual-path attention-fusion convolutional neural network for multimodal brain tumor segmentation, Biomed. Signal. Process. 79 (2023) 104037.
[15] M. Lin, S. Wang, Y. Ding, L. Zhao, F. Wang, Y. Peng, An empirical study of using radiology reports and images to improve ICU-mortality prediction, in: Proceedings of IEEE International Conference on Healthcare Informatics, 2021, pp. 497–498.
[16] J.Venugopalan L. Tong, H.R. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of Alzheimer's disease stage, Sci. Rep. 11 (2021) 3254.
[17] V. Guarrasi, L. Tronchin, D. Albano, E. Faiella, D. Fazzini, D. Santucci, P. Soda, Multimodal explainability via latent shift applied to COVID-19 stratification, Pattern Recognit. 156 (2024) 110825.
[18] B.V. Amsterdam, I. Funke, E. Edwards, S. Speidel, J. Collins, A. Sridhar, J. Kelly, M.J. Clarkson, D. Stoyanov, Gesture recognition in robotic surgery with multimodal attention, IEEE Trans. Med. Imaging 41 (7) (2022) 1677–1687.
[19] W. Xiao, L. Yuan, T. Ran, L. He, J. Zhang, J. Cui, Multimodal fusion for autonomous navigation via deep reinforcement learning with sparse rewards and hindsight experience replay, Displays 78 (2023) 102440.
[20] Y. Liu, X. Zhang, Q. Zhang, C. Li, F. Huang, X. Tang, Z. Li, Dual self-attention with co-attention networks for visual question answering, Pattern Recognit. 117 (2021) 107956.
[21] Y. Wei, L. Ji, Multi-modal bilinear fusion with hybrid attention mechanism for multi-label skin lesion classifcation, Multimedia Tools Appl. 83 (2024) 65221–65247.
[22] J.Y. Koh, R. Salakhutdinov, D. Fried, Grounding language models to images for multimodal inputs and outputs, in: Proceedings of International Conference on Machine Learning, 2023, pp. 17283–17300.
[23] X. Mei, L. Yang, D. Gao, X. Cai, J. Han, T. Liu, PhraseAug: An augmented medical report generation model with phrasebook, IEEE Trans. Med. Imaging 43 (12) (2024) 4211–4223.
[24] Q. Zhu, H. Wang, B. Xu, Z. Zhang, W. Shao, D. Zhang, Multimodal triplet attention network for brain disease diagnosis, IEEE Trans. Med. Imaging 41 (12) (2022) 3884–3894.
[25] Y. Guo, A mutual attention based multimodal fusion for fake news detection on social network, Appl. Intell. 53 (2023) 15311–15320.
[26] X. Huang, H. Gong, A dual-attention learning network with word and sentence embedding for medical visual question answering, IEEE Trans. Med. Imaging 43 (2) (2024) 832–845.
[27] 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, in: Proceedings of International Conference on Machine Learning, Vol. 139, 2021, pp. 8748–87637.
[28] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inform. Theory 37 (1) (1991) 145–151.
[29] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.
[31] J. Mao, S. Qiu, W. Wei, H. He, Cross-modal guiding and reweighting network for multi-modal RSVP-based target detection, Neural Netw. 161 (2023) 65–82.
[32] T. Zhou, H. Fu, G. Chen, Y. Zhou, D. Fan, L. Shao, Specificity-preserving RGB-D saliency detection, in: Proceedings of IEEE International Conference on Computer Vision, 2021, pp. 4661–4671.
[33] M. Wortsman, J. Lee, J. Gilmer, S. Kornblith, Replacing softmax with ReLU in vision transformers, 2023, arXiv:2309.08586.
[34] 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: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[35] Z. Pan, F. Wu, B. Zhang, Fine-grained image-text matching by cross-modal hard aligning network, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19275–19284.