MCD-CLIP: Multi-view Chest Disease Diagnosis with Disentangled CLIP

Songyue Cai[1] , **Yujie Mo**[1*] , **Liang Peng**[2] , **Yucheng Xie**[1] , **Tao Tong**[1] , **Xiaofeng Zhu**[1]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong
sonnycai@std.uestc.edu.cn, {moyujie2017, larrypengliang, xyemrsnon, tongqtao, seanzhuxf}@gmail.com

Abstract

Pre-trained methods for multi-view chest X-ray images have demonstrated impressive performance in chest disease diagnosis, but there are still some limitations that need to be addressed. Firstly, many pre-trained methods require full fine-tuning pre-trained models to induce significant computational resource usage and the prior knowledge destruction. Secondly, many pre-trained methods cannot efficiently balance consistency and complementarity among views, leading to information loss and performance degradation. To tackle these issues, we propose MCD-CLIP, a CLIP-based multi-view chest disease diagnosis method. It uses visual prompts and the Prompt-Aligner to align prompts across views, along with the additional text representation for efficient transfer. Moreover, we employ adapters to disentangle the image representation, maintaining consistency and complementarity from different views. Experimental results on the chest X-ray dataset demonstrate that MCD-CLIP achieves comparable or better performance on a variety of tasks with 94.31% fewer tunable parameters compared to state-of-the-art methods. The source codes are released at https://github.com/YuzunoKawori/MCD-CLIP.

1 Introduction

Multi-view (or multi-modality) chest X-ray images are crucial in diagnosing chest disease, as they can capture more comprehensive information compared to single-view data [Raoof *et al.*, 2012]. For instance, the lateral view of a chest X-ray image reveals areas of lungs that are not visible in the frontal view [Feigin, 2010]. Hence, recent deep learning methods for computer-aided diagnosis of chest disease pay more attention on multi-view data [Qin *et al.*, 2018].

Previous deep learning methods for chest disease diagnosis using multi-view data can be classified into two categories, *i.e.,* traditional deep learning methods and pre-trained methods. Traditional deep learning methods use deep neural networks as feature encoders to extract features from multi-view chest X-ray images. For instance, DualNet [Rubin *et al.*, 2018] trains a modified version of DenseNet-121 [Huang *et al.*, 2017] to extract features for each view independently, and then concatenates features from different views. MVCNET [Zhu and Feng, 2021] incorporates a BPT-branch to fuse features between two views during the feature extraction stage. However, traditional deep learning methods are difficult to extract the intrinsic connection of features cross different views, so that they cannot effectively utilize multi-view information.

Recently, pre-trained methods increasingly leverage the attention mechanism of transformer [Vaswani, 2017] to either discover local correlations among views or capture view-specific information. For instance, CVT [Van Tulder *et al.*, 2021] introduces attention blocks of the cross-view transformer to fuse feature maps generated by ResNet-18 [He *et al.*, 2016], thereby capturing local correlations across views. MV-HFMD [Black and Souvenir, 2024] employs knowledge mutual distillation to merge complementarity extracted by transformer from different views. Chest X-ray images often have multiple views and are limited by the sample size, which can lead to over-fitting. However, pre-trained models can effectively mitigate over-fitting due to their strong generalization ability, which is why they are gaining increasing attention in the field of chest disease diagnosis.

However, previous pre-trained methods still have limitations that need to be addressed. Firstly, existing methods generally use the full fine-tuning to adapt pre-trained models to downstream tasks for effectiveness. However, the full fine-tuning paradigm often destroys the prior knowledge of pre-trained models, as well as results in the degradation of the model generalization ability. For instance, while CVT excels at fusing information from different views during feature extraction, it is typically limited to dual-view data and is difficult to extend to multiple views. Secondly, existing methods generally cannot efficiently balance consistency and complementarity among views, despite their importance for downstream tasks in various domains [Xie *et al.*, 2020; Wang *et al.*, 2022; Mo *et al.*, 2023]. As a result, this may result in either performance degradation or lead to excessive computational costs. For instance, MV-HFMD employs transformers with unshared weights, increasing the model's computational burden and potentially introducing redundant

*Corresponding Author

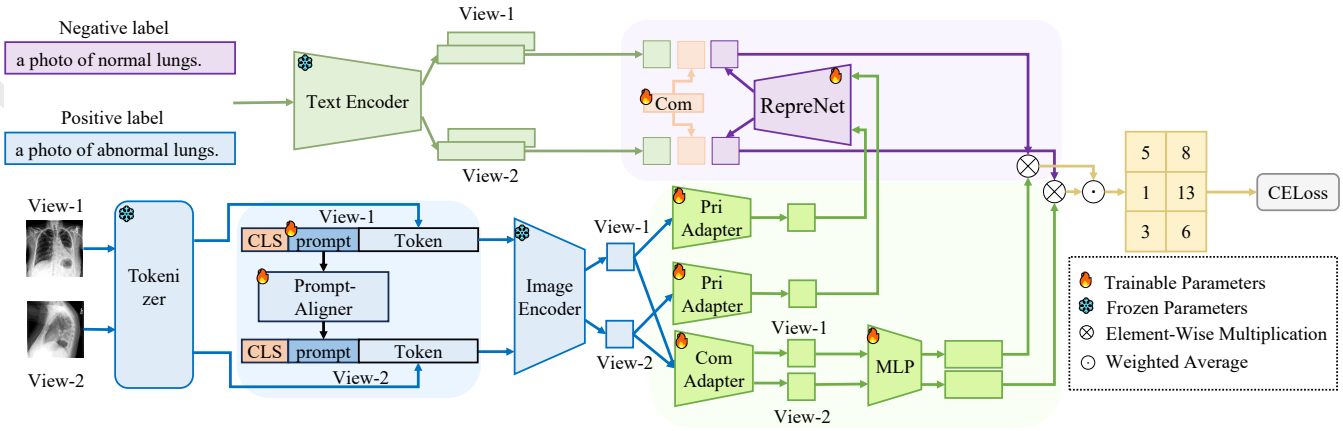

Figure 1: The flowchart of the proposed MCD-CLIP, involving three modules, *i.e.,* the Prompt Alignment (blue block), the Representation Disentanglement (green block), and the Text Representation Enhancement (purple block). Specifically, given a multi-view image, MCD-CLIP first obtains its token embedding using the CLIP tokenizer, then adds visual prompts for the first view. The Prompt Alignment is further designed to obtain the visual prompts for the remaining views. Next, MCD-CLIP obtains the image representation using the frozen pre-trained image encoder. The Representation Disentanglement disentangles the image representation for each view into the common image and the private image representation. The common image representation is used as the final image representation, which is projected to the same dimension as the final text representation via a Multi-Layer Perceptron (MLP). For the text modality, MCD-CLIP uses a fixed template that is encoded into the text representation by the frozen pre-trained text encoder. Additionally, in the Text Representation Enhancement module, it adds the common text representation with learnable parameters and the private text representation, which obtained from the private image representation by RepreNet to the text representation as the final text representation. Finally, the similarities cross different views are fused using weighted averaging to generate the final decision.

information.

To address the above issues, in this paper, we propose a new method with the Prompt Alignment, the Representation Disentanglement, and the Text Representation Enhancement for Multi-view Chest Disease Diagnosis with the pre-trained CLIP (MCD-CLIP for short), as shown in Figure 1. Specifically, in the Prompt Alignment, we first introduce visual prompts to adapt pre-trained models to medical image domain. After that, we investigate the Prompt-Aligner to align visual prompts of different views. As a result, the Prompt Alignment efficiently tunes prompts with few parameters maintaining prompts consistency across views, thus exploring the first issue of previous methods. In the Representation Disentanglement, we design the common adapter and the private adapter to disentangle the image representation from the frozen pre-trained image encoder into the common image representation and the private image representation. As a result, the Representation Disentanglement efficiently maintains consistency cross views and complementarity in each view, thus exploring the second issue of previous methods. In the Text Representation Enhancement, we further add the additional text representation to enhance the expressiveness of the text representation. As a result, the Text Representation Enhancement efficiently adapts CLIP [Radford *et al.*, 2021] to multi-view chest X-ray images. Compared to previous methods, the main contributions of our method can be summarized as:

- We propose a new multi-view efficient transfer method, which can be adapted to the diagnosis of chest disease by training only a small number of parameters while improving the expressiveness of the text representation by adding the additional text representation.
- We propose a new fusion method to efficiently extract the representation of multi-view data by preserving both consistency and complementarity across views.
- We demonstrate the effectiveness of our method on the large public dataset with five scenarios applied to the sample labels. We are the first work to conduct all scenarios applied to the sample labels in one framework on multi-view chest disease diagnosis.

2 Method

2.1 Motivation

Existing pre-trained methods generally adapt to multi-view chest disease diagnosis by modifying pre-trained parameters in a full fine-tuning strategy [Van Tulder *et al.*, 2021]. However, the full fine-tuning strategy updates all parameters of pre-trained models, which requires a lot of matrix computations and backpropagation steps, leading to expensive computational resources. Moreover, the full fine-tuning strategy destroys the prior knowledge in pre-trained models and may cause over-fitting to task-specific data, leading to suboptimal generalization ability of pre-trained models. Therefore, it is necessary to adopt efficient transfer learning to maintain good generalization performance of pre-trained models.

In addition, chest X-ray images usually contain multiple views, describing the specific lesion area from different viewpoints. Therefore, different views contain consistency of the lesion area, accurately describing the lesion area [Hashir *et al.*, 2020]. Moreover, each view contains complementarity

to other views, providing a more comprehensive description of the lesion area [Ittyachen *et al.*, 2017]. To obtain consistency or complementarity from multiple views, existing methods for multi-view chest disease diagnosis usually use middle, late, and hybrid fusion mechanisms [Black and Souvenir, 2024]. However, all of these mechanisms cannot efficiently and effectively balance consistency and complementarity among views well. Specifically, the middle and late fusion can maintain either consistency or complementarity only, leading to a decline in another. In addition, the hybrid fusion balances consistency and complementarity among different views, but it incurs the cost of computing resources. Therefore, it is necessary to efficiently consider both consistency and complementarity among views.

Based on the above analysis, it is intuitive to tune pre-trained models as well as capture consistency and complementarity in an efficient and effective way. Therefore, there are two questions to be answered. First, how to achieve multi-view disease diagnosis by tune few parameters to improve the model efficiency? Second, how to simultaneously maintain consistency and complementarity efficiently and effectively, under their conflict nature?

However, existing methods rarely consider these issues comprehensively. Recently, surprising progress has been made in efficient transfer learning based on CLIP [Radford *et al.*, 2021], migrating pre-trained models to downstream tasks with few tunable parameters [Zhou *et al.*, 2022a; Zhou *et al.*, 2022b; Ghosal *et al.*, 2024; Gao *et al.*, 2024a; Chen *et al.*, 2024; Zhang *et al.*, 2022]. Unfortunately, due to the multi-view nature of X-ray chest images, direct application of these methods for single-view data may result in poor performance degradation. In addition, many medical applications, such as synthesis [Ben-Cohen *et al.*, 2019], cross modality segmentation [Yang *et al.*, 2019], and generating rib-suppressed chest X-ray [Han *et al.*, 2022], adopt the concept of representation disentanglement to extract distinct types of representation, thereby enhancing performance. However, they may not be suitable for chest disease diagnosis. These applications are limited by performance of encoders and do not achieve semantic disentanglement.

Therefore, in this paper, we propose a novel multi-view chest disease diagnosis framework with the Prompt Alignment, the Representation Disentanglement, and the Text Representation Enhancement to address the two issues mentioned above.

2.2 Preliminary

Our method designs the Prompt Alignment, the Representation Disentanglement, and the Text Representation Enhancement to adapt pre-trained models to medical image data. To do this, we first briefly introduce the widely adopted pre-trained model, *i.e.,* Contrastive Language-Image Pre-training (CLIP) [Radford *et al.*, 2021]. Specifically, it consists of two modules, *i.e.,* image encoder and text encoder:

(1) Image encoder $\mathcal{V}(\cdot)$ aims to convert image into the image presentation through tokenization and encoding operations. Specifically, the tokenization divides the original image into a set of patches $\boldsymbol{H} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_n}\}$ and embedded in patches to obtain a set of token embedding $\boldsymbol{T} = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \cdots, \boldsymbol{t_n}\}$. After that, the encoding employs the ViT [Dosovitskiy, 2020] structure to obtain the final potential image presentation $\boldsymbol{z} \in \mathbb{R}^{1\times D}$, *i.e.,*

$$\boldsymbol{e_0} = [\boldsymbol{x_{cls}}, \boldsymbol{t_1}, \boldsymbol{t_2}, \cdots, \boldsymbol{t_n}] + \boldsymbol{E_{pos}}, \tag{1}$$

$$\boldsymbol{z} = \text{ViT}(\boldsymbol{e_0}), \tag{2}$$

where $\boldsymbol{E_{pos}} \in \mathbb{R}^{(n+1)\times D}$ is the positional embedding, $\text{ViT}(\cdot)$ is the frozen pre-trained image encoder, $\boldsymbol{x_{cls}}$ is pre-trained class token, n is the number of patches and D is the dimension of the final embedded features, $\boldsymbol{z}$ is obtained by applying a projection layer to the $\boldsymbol{x_{cls}}$ of the last transformer layer.

(2) Text encoder $\mathcal{T}(\cdot)$ aims to convert text prompts into the text representation. Similar to the image encoder, the text encoder also includes the tokenization and encoding operations. To do this, CLIP first generates text prompts manually. Specifically, for tasks with c-class classification, its category label is $CLS = \{cls_1, cls_2, \cdots, cls_c\}$, text prompts for the i-th class can be represented as "a photo of a $[cls_i]$." After that, the tokenization divides text prompts into a set of token embedding $\boldsymbol{M} = \{\boldsymbol{m_1}, \boldsymbol{m_2}, \cdots, \boldsymbol{m_c}\}$. In addition, the encoding also employs a transformer-based model [Vaswani, 2017] to obtain the text representation, *i.e.,*

$$\boldsymbol{W} = \mathcal{T}(\boldsymbol{M}),\ \boldsymbol{W} \in \mathbb{R}^{c\times D}, \tag{3}$$

where c is number of classes.

Finally, for a given image $\mathcal{I}$, its probability to each target class can be given by:

$$p\left(y=i|\boldsymbol{\mathcal{I}}\right) = \frac{\exp\left(\cos\left(\mathcal{V}\left(\boldsymbol{\mathcal{I}}\right), \mathcal{T}\left(\boldsymbol{m_i}\right)\right)/\tau\right)}{\sum_{j=1}^{c}\exp\left(\cos\left(\mathcal{V}\left(\boldsymbol{\mathcal{I}}\right), \mathcal{T}\left(\boldsymbol{m_j}\right)\right)/\tau\right)}, \tag{4}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity operation, τ is the temperature coefficient. The class token within each prompt $\boldsymbol{m_i}$ is replaced by the corresponding word embedding vectors of the i-th class name.

2.3 Prompt Alignment

Previous methods use the full fine-tuning to adapt pre-trained model parameters to multi-view chest X-ray datasets [Black and Souvenir, 2024; Van Tulder *et al.*, 2021]. However, this not only destroys the prior knowledge in pre-trained models but also weakens generalization and scalability of pre-trained models. Therefore, to address this issue, a intuitive idea is to investigate efficient transfer learning to achieve tuning of pre-trained models with a small number of parameters.

To do this, we refer to existing prompt-tuning methods (*e.g.,* VPT [Jia *et al.*, 2022]) for pre-trained models by tuning a few learnable prompts. Specifically, we first insert a set of learnable shallow prompts between the learnable class token and the token embedding. After that, we freeze encoder backbone network parameters and only tune prompts during training, *i.e.,*

$$\boldsymbol{z_0} = [\boldsymbol{x_{cls}}, \boldsymbol{P}, \boldsymbol{T'}], \boldsymbol{z_0} \in \mathbb{R}^{(n+k+1)\times D}, \tag{5}$$

where $\boldsymbol{P} \in \mathbb{R}^{k\times D}$ is a set of learnable prompts, $\boldsymbol{T'} \in \mathbb{R}^{n\times D}$ is obtained by summing the token embedding and the position embedding, and k is the number of prompts. As a result,

such prompt-tuning adapts the model to understand different downstream tasks with few parameters.

However, the above prompt-tuning aims to optimize pre-trained models for single-view data. As a result, pre-trained models may ignore prompts consistency of different views of the same sample when inputs are multi-view data. Therefore, unconstrained prompts may prevent the model from recognizing that the views originate from the same sample, ultimately degrading performance of multi-view classification.

Based on the above analysis, we hope pre-trained models can understand prompts consistency of different views from the same sample, thus outputting discriminative representation for downstream tasks. To do this, a natural idea is to align prompts from different views to capture consistency. Therefore, inspired by the multi-modal pre-trained model MaPLe [Khattak *et al.*, 2023], we investigate a Prompt-Aligner to align prompts from different views. Specifically, while prompts for the first view are directly generated using tunable parameters, prompts for the other views are derived by the Prompt-Aligner, *i.e.,*

$$\boldsymbol{P_i} = \text{ProAligner}_i(\boldsymbol{P_1}), \quad i \geq 2, \tag{6}$$

where $\text{ProAligner}_i(\cdot)$ is a linear layer to align prompts. As a result, we use the Prompt-Aligner to map prompts of different views into the similar semantic space, which maintains prompts consistency of different views.

Compared to the original VPT method, our method is able to maintain prompts consistency of different views. Moreover, our method does not introduce extra tunable parameters than the VPT method, as our method only trains the parameters of the initial view's prompt and the Prompt-Aligner, thus achieving efficiency.

2.4 Multi-View Representation Disentanglement

With the Prompt Alignment, we input multi-view chest X-ray data along with corresponding prompts into the frozen pre-trained image encoder, resulting in the image representation that capture discriminative information in each view. Actually, the image representation of different views reflect characteristics of the sample as observed from various perspectives. As a result, they reflect consistency of the sample, but also contain complementarity that the other views do not have. However, directly utilizing the image representation from the frozen pre-trained image encoder may mix consistency and complementarity to cause negative impacts [Yang *et al.*, 2022]. Therefore, it is crucial to effectively fuse information from different views. Unfortunately, existing pre-trained methods struggle to balance consistency and complementarity effectively [Van Tulder *et al.*, 2021] or result in high computational costs [Black and Souvenir, 2024]. Consequently, these methods not only hinder model transferability but also impose significant computational overhead.

To address these challenges, we propose to disentangle the image representation extracted by the frozen pre-trained image encoder. Specifically, we disentangle the image representation from different views into the common image representation and the private image representation.

First, we design the Common-Adapter to map the image representation from different views to the same common semantic space. Formally, for the i-th view, its common representation is obtained as follows:

$$\boldsymbol{z^i_{com}} = \text{ComAdapter}(\boldsymbol{z^i}), \quad \boldsymbol{z^i_{com}} \in \mathbb{R}^{1\times D}, \tag{7}$$

where all views use the same weighted $\text{ComAdapter}(\cdot)$. Moreover, we design the Common-Adapter by following the structure of the CLIP-Adapter [Gao *et al.*, 2024b], which is built with a two-layer neural network structure. For the common image representation of different views, they are inclined to depict the common characteristics that observed from different views, as the same subject will present different features under different viewpoints. As a result, our method designs a Common-Adapter with shared weights to better extract consistency among views.

After that, we further design different Private-Adapters to obtain the private image representation of different views. Formally, for the i-th view, the private image representation is obtained as follows:

$$\boldsymbol{z^i_{pri}} = \text{PriAdapter}_i(\boldsymbol{z^i}), \quad \boldsymbol{z^i_{pri}}, \boldsymbol{z^i} \in \mathbb{R}^{1\times D}, \tag{8}$$

where $\boldsymbol{z^i}$ is the image representation of the i-th view. For the private representation of different views, they are expected to contain different information from different viewpoints. As a result, our method designs separate Private-Adapters with non-shared weights to better extract complementarity that is unique to each view.

Finally, in this module, we train only a few parameters of Adapters to disentangle the image representation, thereby obtaining both the common image representation and the private image representation for different views. As a result, the common image representation highlights the information shared among views, ensuring consistency among them. Meanwhile, the private image representation retains unique information specific to each view as much as possible, ensuring complementarity in each view. Overall, the module realizes the balance between consistency and complementarity in multi-view fusion with only a small number of parameters, thus further achieving the efficiency.

2.5 Disentangled Representation Utilization

With the Representation Disentanglement, our method divides the image representation into the common image representation and the private image representation. As a result, the common image representation contains consistency, but cannot capture more detailed information in each view. Moreover, the private image representation contains complementarity from each view, but it cannot accurately describe the main information among views. Therefore, it is a challenge to use the common image representation and the private image representation effectively, as they cannot be directly fused for decision-making due to their distinct characteristics. To do this, we use the private image representation to enhance the expressiveness of the text representation. Moreover, we use the common image representation for the multi-view classification decision.

Text Representation Enhancement

Under the setting of text prompts, one popular method uses a fixed template that manually selected, *e.g.,* "a photo of a

$[cls]$". The fixed manual template has a good generalization performance, but the quality of the fixed template may have a large impact to performance [Zhou *et al.*, 2022b]. Moreover, only using a fixed template with no tunable parameters cannot understand the specialized medical terminology well. As a result, semantic bias may occur and result in performance degradation. To alleviate the issue of the manually fixed template, another popular practice is CoOp [Zhou *et al.*, 2022b], which uses optimized parameters to automatically select the template and shows good performance on most tasks. However, CoOp accidentally memorizes background information of the target class during the training phase [Ming *et al.*, 2022; Miyai *et al.*, 2023]. Moreover, most of the backgrounds of the medical images are often similar and meaningless, memorizing these background information may accidentally reduce the performance of the model.

To alleviate the above issues, we propose a fixed template enhancement technique to enrich the expressiveness of the text representation by adding the additional text representation, *i.e.,* the common text representation and the private text representation. Specifically, we first input the fixed template prompts (*e.g.,* "a photo of normal/ abnormal lungs.") to the frozen pre-trained text encoder thus obtain the text representation $\boldsymbol{W} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_c}\}$, and then propose a learnable common text representation $\boldsymbol{r_{com}}$, followed by adding it to the text representation to capture and describe consistency among views. Formally, the common text representation is:

$$\boldsymbol{w'_j} = [\boldsymbol{w_j}, \boldsymbol{r_{com}}], \quad \boldsymbol{w'_j} \in \mathbb{R}^{1\times(A+D)}, \boldsymbol{r_{com}} \in \mathbb{R}^{1\times A}, \tag{9}$$

where A is the dimension of the common text representation. In Eq. (9), the common text representation complements the fixed text template to provide consistency for multi-view data. As a result, the common text representation enhances the expressiveness of consistency among views.

Moreover, due to differences among views, different views observe different parts and perspectives of chest X-ray data. If the differences among views cannot be distinguished and only the common text representation is used, it may lead to information loss as well as incorrect judgments. Therefore, we utilize the private image representation as the private text representation for different views, because the private image representation contains rich complementarity from different views. However, the semantic space of the text and the image representation are different. To address this issue, we design a RepreNet to further convert complementarity of different views and obtain the private text representation, *i.e.,*

$$\boldsymbol{r^i_{pri}} = \text{RepreNet}(\boldsymbol{z^i_{pri}}), \quad \boldsymbol{r^i_{pri}} \in \mathbb{R}^{1\times B}, \tag{10}$$

where RepreNet$(\cdot)$ is designed as a linear layer or Multi-Layer Perceptron (MLP), and B is the dimension of the private representation. In Eq. (10), the RepreNet converts the private image representation into the private text representation. As a result, the RepreNet improves the quality of the private text representation and achieves the cross-modal fusion.

After that, we add the private text representation to the text representation as the final text representation, thus capturing and describing complementarity from different views, *i.e.,* $\boldsymbol{X^i} = \left\{\boldsymbol{x^i_1}, \boldsymbol{x^i_2}, \cdots, \boldsymbol{x^i_c}\right\}$ for the i-th view, where

$$\boldsymbol{x^i_j} = [\boldsymbol{w'_j}, \boldsymbol{r^i_{pri}}], \quad \boldsymbol{x^i_j} \in \mathbb{R}^{1\times(A+B+D)}. \tag{11}$$

In Eq. (11), the private text representation complements the fixed text template to provide unique information different from other views. As a result, the private text representation enhances the expressiveness of complementarity.

In this module, we generate the common text representation with learnable parameters and obtain the private text representation by fusing the visual modality, so it enriches the expressiveness of the text presentation and preserves the generalization properties of the text fixed template.

Multi-View Fusion

We use the final text representation and the common image representation as the final image representation to compute prediction probabilities of different views for the target class separately. To do this, we first compute the cosine similarity of different views for each class. the cosine similarity between the i-th view and the j-th class can be given by:

$$\boldsymbol{s^i_j} = \cos\left(\text{MLP}(\boldsymbol{z^i_{com}}), \boldsymbol{x^i_j}\right), \tag{12}$$

where MLP$(\cdot)$ is used to map the final image representation to the same dimensions as the final text representation.

However, the quality of the representation from different views is different. For instance, for multi-view X-ray chest image, the lateral view contains the information which is not present in the frontal view, enriching the feature expressiveness of the sample. However, the frontal view still observes most of the chest area. Therefore, for poor quality lateral views with less information, we propose to reduce influence of those views in the decision-making stage. Moreover, for the high-quality frontal views, we propose to increase their influence. Based on the above analysis, we assign weights of cosine similarities of different views, *i.e.,* the similarity of the i-th class after weighted fusion for two views is:

$$\boldsymbol{s_i} = \alpha \times \boldsymbol{s^1_i} + (1-\alpha) \times \boldsymbol{s^2_i}, \tag{13}$$

where α is the view weighting factor. Moreover, it allows the selection of appropriate weighting ratios under different tasks. As a result, our method enhances generalization of the model across different tasks by weighting and fusing the similarity of different views.

Finally, for an multi-view chest X-ray data input, its probability to the i-th class can be given by:

$$p\left(y=i|\{\boldsymbol{\mathcal{I}_1}, \boldsymbol{\mathcal{I}_2}\}\right) = \frac{\exp\left(\boldsymbol{s_i}/\tau\right)}{\sum_{j=1}^{c}\exp\left(\boldsymbol{s_j}/\tau\right)}. \tag{14}$$

Moreover, during the model training, all tunable parameters are optimized by cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{c} y_i^n \log \hat{y}_i^n, \tag{15}$$

where N is the total number of training examples, $y_i = 1$ if it equals to the truth category, otherwise $y_i = 0$, and the value of the $\hat{y}_i$ equals to $p\left(y=i|\{\boldsymbol{\mathcal{I}_1}, \boldsymbol{\mathcal{I}_2}\}\right)$.

Task	PointCLIP	TaskRes	CLIP-Adapter	CoOp	Lin.CLIP	Zer.CLIP	DualNet	MVC-NET	CVT	MV-HFMD	Ours
Atelectasis	0.700	0.574	0.553	0.718	0.634	0.529	0.663	0.650	0.716	0.745	**0.759**
Cardiomegaly	0.844	0.568	0.624	0.847	0.791	0.578	**0.884**	0.878	0.862	0.871	0.872
Consolidation	0.786	0.517	0.579	0.794	0.711	0.532	0.802	0.790	0.809	0.821	**0.825**
Edema	0.825	0.467	0.552	0.822	0.771	0.520	0.812	0.807	0.841	**0.854**	0.843
Enlarged Cardiomed.	0.753	0.481	0.583	0.757	0.711	0.558	0.760	0.747	0.761	0.766	**0.776**
Fracture	0.716	0.508	0.521	0.717	0.624	0.518	0.701	0.679	0.701	0.745	**0.765**
Lung Lesion	0.679	0.482	0.526	**0.713**	0.639	0.514	0.628	0.638	0.679	0.693	0.700
Lung Opacity	0.739	0.495	0.599	0.735	0.677	0.561	0.730	0.712	0.736	0.722	**0.746**
Pleural Effusion	0.891	0.591	0.663	0.874	0.813	0.547	0.906	0.904	0.917	0.921	**0.923**
Pleural Other	0.658	0.543	0.617	0.691	0.661	0.591	0.657	0.669	0.742	0.751	**0.773**
Pneumonia	0.689	0.551	0.569	0.681	0.623	0.542	0.665	0.687	0.701	0.704	**0.723**
Pneumothorax	0.787	0.456	0.617	0.765	0.701	0.586	0.776	0.762	0.805	0.810	**0.810**
Support Devices	0.664	0.549	0.531	0.672	0.625	0.563	0.607	0.625	**0.702**	0.676	0.700
Average	0.748	0.522	0.579	0.753	0.691	0.549	0.738	0.735	0.767	0.775	**0.786**

Table 1: The average AUC-ROC of all methods with five scenarios on CheXpert. "Average" is the average results over 13 classification tasks.

3 Experiments

3.1 Experimental settings

Dataset

CheXpert [Irvin *et al.*, 2019] is a large publicly available chest X-ray image dataset. We follow [Van Tulder *et al.*, 2021; Black and Souvenir, 2024] to filter, pre-process, and divide the original dataset. CheXpert includes four kinds of labels, *i.e.,* "positive", "negative", "uncertain" , and "unknown". Previous studies regard the "uncertain" label as one of five scenarios, *i.e.,* U-Ignore, U-Zeros, U-Ones, U-SelfTrained and U-MultiClass. However, to the best of our knowledge, there is no study considering all five scenarios on multi-view chest disease diagnosis. Therefore, to provide a more comprehensive and in-depth evaluation of our method, we report the average results using all five scenarios.

Comparison Methods

The comparison methods include CLIP-based methods, traditional methods, and pre-trained methods. Specifically, the comparison methods include six CLIP-based methods PointCLIP [Zhang *et al.*, 2022], TaskRes [Yu *et al.*, 2023], CLIP-Adapter [Gao *et al.*, 2024b], CoOp [Zhou *et al.*, 2022b], Linear-probe CLIP [Radford *et al.*, 2021], and Zero-shot CLIP [Radford *et al.*, 2021], two traditional deep learning methods DualNet [Rubin *et al.*, 2018] and MVC-NET [Zhu and Feng, 2021], two pre-trained methods CVT [Van Tulder *et al.*, 2021] and MV-HFMD [Black and Souvenir, 2024].

Implementation Details

All experiments are conducted with 2 NVIDIA GeForce RTX-4090 GPUs. We use the CLIP pre-trained on ImageNet [Deng *et al.*, 2009] as the backbone model. For all CLIP-based methods, we uniformly use ViT-B/32 as the image encoder. Unless otherwise specified, all methods are optimized using stochastic gradient descent with a fixed learning rate of 0.0001, weight decay of 10e-5, a view weighting factor α of 0.5, a batch size of 64.

3.2 Experimental Results

Performance Comparison

We evaluate the effectiveness of our method on a large-scale multi-view chest X-ray dataset. Table 1 summarizes the average performance of all methods. Obviously, our method achieves competitive performance than previous methods.

Method	Tunable Param (M)	AUC-ROC
DualNet [Rubin *et al.*, 2018]	47.02	0.738
MVC-NET [Zhu and Feng, 2021]	71.10	0.735
CVT [Van Tulder *et al.*, 2021]	23.67	0.767
MV-HFMD [Black and Souvenir, 2024]	36.05	0.775
Ours	**2.05**	**0.786**

Table 2: Comparison of the number of tunable parameters and the AUC-ROC for different methods on CheXpert.

Firstly, experimental results demonstrate that our method achieves better performance than traditional deep learning methods. For example, it shows an average performance improvement of 6.5% compared to the best traditional deep learning method (*i.e.,* DualNet). This indicates that pre-trained models have rich semantic space and reduced data dependency. As a result, pre-trained models are able to explore correlations among views more effectively, thus better learning discriminative representation.

Secondly, experimental results demonstrate that our method achieves competitive performance than pre-trained methods. For instance, it achieves an average performance improvement of 1.42% compared to the best competitor (*i.e.,* MV-HFMD). This can be attributed to the fact that our method effectively disentangles the representation and ensures consistency and complementarity from different views. As a result, our method efficiently leverages diverse representation, leading to high performance on downstream tasks.

Thirdly, experimental results demonstrate that our method outperforms CLIP-based methods. For instance, it shows an improvement of 5.08% compared to the CLIP-based multi-view classification method (*i.e.,* PointCLIP). A key factor behind this improvement is the design of a specialized efficient transfer method tailored for multi-view data, along with the inclusion of the additional text representation. As a result, our method enhances both consistency of visual prompts and the expressiveness of the text representation.

Ablation Study

Our method includes three key components, *i.e.,* the Prompt Alignment to align visual prompts from each view, the Rep-

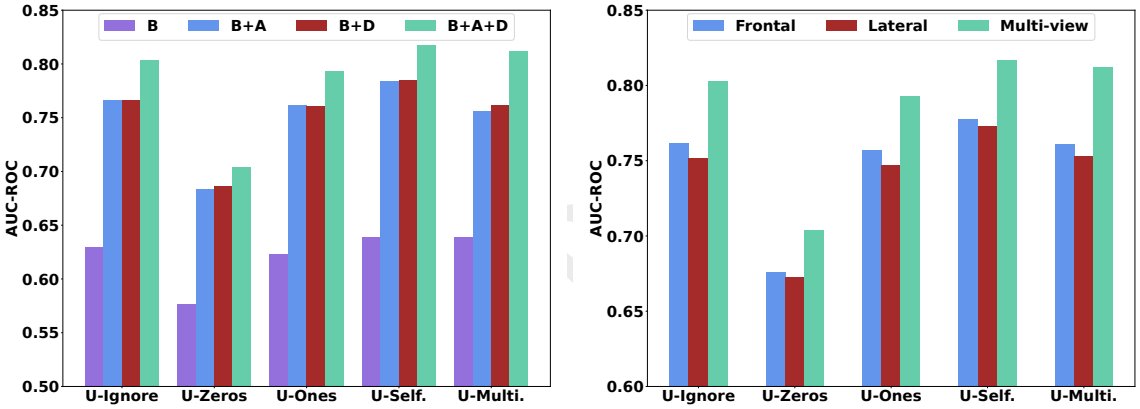

Figure 2: The average AUC-ROC of our method with five scenarios at different combinations of components (left) and different views (right) on CheXpert.

resentation Disentanglement to disentangle the common and private image representation, and the Text Representation Enhancement to add the additional text representation for each view. To validate effectiveness of each component, we evaluate performance of all variants across 13 tasks, with the results presented in Figure 2. However, we do not conduct a separate ablation study on the Text Representation Enhancement, as the private text representation is derived from the Representation Disentanglement. Specifically, we first generate independent visual prompts for each view as the baseline method 'B', apply the Prompt Alignment ('B+A') as well as apply the Representation Disentanglement and the Text Representation Enhancement ('B+D'), and combining all three components ('B+A+D') as our method.

Based on the experimental results, we observe that each component contributes significantly to the overall performance. First, the addition of the Prompt Alignment improves performance by 20.77% compared to the baseline method, demonstrating effectiveness of efficient transfer learning and ensuring consistency in visual prompts across views for multi-view chest X-ray images. Moreover, the Representation Disentanglement and the Text Representation Enhancement improve performance by 21.10% compared to the baseline method. The Representation Disentanglement effectively captures consistency and complementarity from different views, while the Text Representation Enhancement enhances the expressiveness of the text representation in medical image domain. Finally, the proposed MCD-CLIP method improve performance by 26.57% compared to the baseline method, underscoring the importance of all components for the success of our method.

In addition, our method utilizes rich information of different views to improve the performance of multi-view chest disease diagnosis. We further investigate effectiveness of single-view and multi-view by reporting performance when utilizing different views in Figure 2. Specifically, the multi-view approach yields a large improvement in performance compared to the direct use of single view classification. For example, multi-view performance is improved by 5.22% compared to the frontal view and 6.22% compared to the lateral view. it demonstrates that utilizing complementary information among views improves performance of the model when diagnosing chest disease.

Efficiency Analysis

Our method tunes pre-trained models as well as efficiently keep the trade-off between consistency and complementarity cross views by a small number of parameters. We report the number of tunable parameters for multi-view chest disease diagnosis methods in Table 2. Specifically, our method uses the efficient transfer learning to tune pre-trained models, and also trains only a few parameters of the adapter to balance consistency and complementarity efficiently among views. For instance, Compared with the best competitor (*i.e.,* MV-HFMD), our method reduces the number of tunable parameters by 94.31% while maintaining comparable performance. As a result, our method performs comparable or better performance with fewer parameters than previous multi-view chest disease diagnosis methods. It is demonstrated that our method achieves efficient multi-view chest disease diagnosis.

4 Conclusion

In this paper, we propose a new CLIP-based multi-view chest disease diagnosis method. Specifically, we design the Prompt Alignment to efficiently tune pre-trained models and contain prompts consistency of different views. In addition, we employ the Representation Disentanglement to disentangle the image representation into the common image representation and the private image representation of each view. As a result, the Representation Disentanglement efficiently preserves consistency and complementarity from different views. Further, we design the Text Representation Enhancement to add the common text representation and the private text representation to the text representation. As a result, it improves the expressiveness of the text representation. Finally, experiments demonstrate that the proposed MCD-CLIP achieves better or comparable performance to the state-of-the-art methods with a small number of parameters.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported in part by the National Key Research & Development Program of China under Grant 2022YFA1004100, the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (ZYGX2022YGRH009 and ZYGX2022YGRH014).

References

[Ben-Cohen *et al.*, 2019] Avi Ben-Cohen, Roey Mechrez, Noa Yedidia, and Hayit Greenspan. Improving cnn training using disentanglement for liver lesion classification in ct. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 886–889, 2019.

[Black and Souvenir, 2024] Samuel Black and Richard Souvenir. Multi-view classification using hybrid fusion and mutual distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 270–280, 2024.

[Chen *et al.*, 2024] Xuxin Chen, Yuheng Li, Mingzhe Hu, Ella Salari, Xiaoqian Chen, Richard LJ Qiu, Bin Zheng, and Xiaofeng Yang. Mammo-clip: Leveraging contrastive language-image pre-training (clip) for enhanced breast cancer diagnosis with multi-view mammography. *arXiv preprint arXiv:2404.15946*, 2024.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Feigin, 2010] David S Feigin. Lateral chest radiograph: a systematic approach. *Academic Radiology*, 17:1560–1566, 2010.

[Gao *et al.*, 2024a] Jingsheng Gao, Jiacheng Ruan, Suncheng Xiang, Zefang Yu, Ke Ji, Mingye Xie, Ting Liu, and Yuzhuo Fu. Lamm: Label alignment for multi-modal prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1815–1823, 2024.

[Gao *et al.*, 2024b] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132:581–595, 2024.

[Ghosal *et al.*, 2024] Soumya Suvra Ghosal, Samyadeep Basu, Soheil Feizi, and Dinesh Manocha. Intcoop: Interpretability-aware vision-language prompt tuning. *arXiv preprint arXiv:2406.13683*, 2024.

[Han *et al.*, 2022] Luyi Han, Yuanyuan Lyu, Cheng Peng, and S Kevin Zhou. Gan-based disentanglement learning for chest x-ray rib suppression. *Medical Image Analysis*, 77:102369, 2022.

[Hashir *et al.*, 2020] Mohammad Hashir, Hadrien Bertrand, and Joseph Paul Cohen. Quantifying the value of lateral views in deep learning for chest x-rays. In *Medical Imaging with Deep Learning*, pages 288–303, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597, 2019.

[Ittyachen *et al.*, 2017] Abraham M Ittyachen, Anuroopa Vijayan, and Megha Isac. The forgotten view: Chest x-ray-lateral view. *Respiratory medicine case reports*, 22:257–259, 2017.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727, 2022.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[Ming *et al.*, 2022] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10051–10059, 2022.

[Miyai *et al.*, 2023] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Advances in Neural Information Processing Systems*, pages 76298–76310, 2023.

[Mo *et al.*, 2023] Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Disentangled multiplex graph representation learning. In *International Conference on Machine Learning*, pages 24983–25005, 2023.

[Qin *et al.*, 2018] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical Engineering Online*, 17:1–23, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[Raoof *et al.*, 2012] Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141:545–558, 2012.

[Rubin *et al.*, 2018] Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

[Van Tulder *et al.*, 2021] Gijs Van Tulder, Yao Tong, and Elena Marchiori. Multi-view analysis of unregistered medical images using cross-view transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 104–113, 2021.

[Vaswani, 2017] A Vaswani. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 6000–6010, 2017.

[Wang *et al.*, 2022] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

[Xie *et al.*, 2020] De Xie, Cheng Deng, Chao Li, Xianglong Liu, and Dacheng Tao. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 29:3626–3637, 2020.

[Yang *et al.*, 2019] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263, 2019.

[Yang *et al.*, 2022] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *The ACM International Conference on Multimedia*, pages 1642–1651, 2022.

[Yu *et al.*, 2023] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.

[Zhang *et al.*, 2022] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022.

[Zhu and Feng, 2021] Xiongfeng Zhu and Qianjin Feng. Mvc-net: Multi-view chest radiograph classification network with deep fusion. In *IEEE International Symposium on Biomedical Imaging*, pages 554–558, 2021.