Rethinking Medical VQA Models: Towards Data-Efficient Learning

Shuning He[1], Da Ren[2], Haiwei Pan[1,*], Kejia Zhang[1], Qing Li[2]
[1]Department of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[*]Corresponding authors. Email: panhaiwei@hrbeu.edu.cn

Abstract

Medical Visual Question Answering (VQA) is a key medical AI challenge, but scarce data limits progress. Current methods prioritize more pre-training data, overlooking medical data's inherent constraints (ethics, privacy, specialization) which cause slower accumulation than public web data. Sole reliance on more data risks rapid performance plateaus. To address these challenges, this paper proposes the Cross-Modality Discriminative Pattern Identification model (CMDPI), which fundamentally rethinks training methodologies to has a data-efficient framework for medical VQA. During pre-training, CMDPI identifies inherent biases in conventional techniques under data scarcity conditions and introduces a co-regularization approach that integrates multiple pretraining techniques for regularization to enhance model generalizability. For fine-tuning, a difference reconstruction mechanism is proposed that effectively preserves unique discriminative features and a head mixup is introduced to further remedy the issue of overfitting. Experimental results demonstrate that CMDPI achieves performance comparable to or surpassing existing methods while requiring substantially fewer pre-training data. Our work shows the viability of optimizing training paradigms rather than pursuing indiscriminate data scaling for advancing medical VQA systems.

***Keywords:** Medical Visual Question Answering, Data-Efficient Learning, Contrastive Learning, Reconstruction.*

1. Introduction

Medical Visual Question Answering (VQA) stands significantly in AI-assisted medical diagnosis, while its development is limited by the lack of high quality data. Existing methods [39, 14] follow the common research line in general domains, which involves pre-training to remedy the lack of data by making use of large-scale data. However, the medical domain differs fundamentally from general-purpose domains in data acquisition. Medical data accumulation is constrained by ethical considerations, privacy regulations, and the need for domain-specific expertise, resulting in a significantly slower growth compared to general data. Consequently, merely scaling pre-training data will inevitably create bottlenecks for advancing medical VQA models in conventional training paradigms.

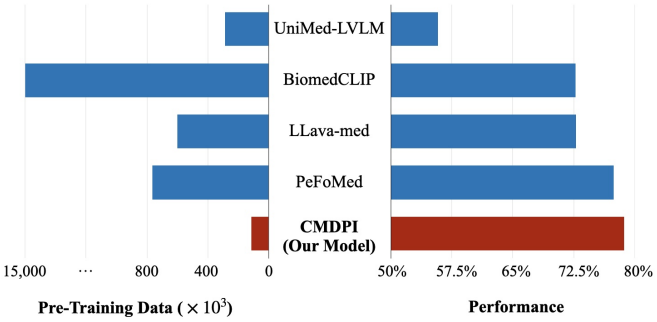


Figure 1: Comparisons of Model Performance on VQA-RAD dataset and Pre-Training Data Size.

This problem reveals the critical need for a systematic investigation of medical VQA models to advance data-efficient learning. Existing models follow two distinct technical approaches: classification-based [35, 17, 16] and generation-based [32, 27, 46]. Classification-based models transform the VQA task into a classification problem by treating each answer as a class label. These methods have flexible structures and thus can be compatible with various techniques. However, their data utilization is limited by the substantial discrepancy between pre-training and fine-tuning tasks. These models often require to adopt a completely new module with randomly initialized parameters for this transformation, preventing the module from benefiting from the pre-training process. Generation-based models, most of which are built upon large language models (LLMs) [23, 52]. These models can fully leverage parameters learned during pre-training and minimal new parameters are introduced during fine-tuning. Nevertheless, their performance on target tasks is highly dependent on dataset size [49], making it challenging to achieve satisfactory results with limited data.

The low data utilization efficiency and high dependence of existing models on extensive pre-training data necessitate a fundamental rethinking of training paradigms. Medical images often exhibit high visual similarities, making the accurate identification of distinctive patterns unique to individual cases crucial for precise question answering. Inspired by this observation, our paper proposes a data-efficient model for medical VQA: the **Cross-Modality Discriminative Pattern Identification Model (CMDPI)**. Its core principle is to enable the identification of discriminative patterns within medical images using limited data. It accomplishes this through two stages—pre-training and fine-tuning—which establish cross-modal alignment and capture discriminative patterns with high data efficiency.

To achieve this objective, CMDPI necessitates a highly data-efficient framework. We therefore redesign existing medical VQA frameworks to integrate the strengths of classification-based and generation-based approaches. This redesigned framework maintains compatibility with existing pre-training techniques, ensuring its hyperparameters transfer seamlessly to fine-tuning. Specifically, modality-specific encoders produce both global and local representations, supporting diverse pre-training objectives. Crucially, answer prediction is unified with the reconstruction process, allowing full leverage of pre-trained parameters without introducing new modules during fine-tuning. This integrated design effectively addresses core limitations of prior models and significantly enhances data utilization efficiency.

For pre-training, currently widely-used techniques such as reconstruction and contrastive learning are predominantly developed for general domains. The effectiveness of these techniques is highly dependent on abundant training data. The reconstruction process enables models to learn data dependencies; however, limited data leads to sparse token distributions, and the dependencies learned by the models become highly random and cannot reflect the true dependencies among data. Similarly, contrastive learning enables models to measure data similarities, yet achieving robust similarity measurements depends on sufficient data to capture representative patterns. When data is limited, these measurements lose generalizability, resulting in inaccurate assessments of data similarities.

To address these problems, our paper leverages the framework's compatibility with diverse pre-training techniques by integrating **co-regularization**. This approach employs multiple pre-training methods to mutually regularize one another, reducing bias induced by limited pre-training data. For reconstruction, we adopt both global reconstruction and local reconstruction. The global reconstruction employs a global representation to guide the reconstruction process, while the local reconstruction, which reconstructs data based on dependencies among local features without global features. The dependencies learned at different granularities can mutually constrain each other and yield more accurate representations. For contrastive learning, both intra- and inter-modality contrastive learning are employed, which leverages similarities within individual modalities to constrain cross-modal representations. Furthermore, our analysis reveals that data type suitability varies across pre-training methods. Consequently, we develop and incorporate a comprehensive CMDPI pre-training approach that adaptively pairs techniques with compatible data to enhance efficiency.

During fine-tuning, models need to perform accurate discriminative pattern identification to answer questions correctly. While models in general domains can accomplish this by leveraging vast amounts of data, such high-quality data is not available on a large scale in the medical domain. To enable models with such capabilities under limited data constraints, a **difference reconstruction** mechanism is proposed. This mechanism requires the model to reconstruct images from similar ones, thereby prompting the model to learn the differences among similar images. By comparing images with similar counterparts, this process effectively guides the model to identify discriminative features under data-limited conditions. To further enhance data efficiency, we extend the idea of Mixup [50] into our framework and adopt a head mixup method to augment data and regularize the model during fine-tuning.

Equipped with the techniques we designed across both pre-training and fine-tuning stages, as shown in Figure 1, CMDPI demonstrates superior performance compared to existing models while requiring substantially less pre-training data. The primary contributions of this work are as follows:

- Our paper conducts a comprehensive analysis of existing methodologies in medical VQA and identify the fundamental limitations inherent in approaches that rely on progressively larger pre-training datasets. To address these challenges and optimize data utilization, a unified framework designed to enhance data efficiency is developed. This framework maximizes the utilization of pre-trained parameters while eliminating the need for randomly initialized parameters during fine-tuning. Furthermore, its modular architecture maintains compatibility with diverse pre-training and fine-tuning techniques, thereby enabling additional performance enhancements.

- Based on this framework, our paper proposes the **Cross-Modality Discriminative Pattern Identification Model (CMDPI)**. We identify fundamental limitations in conventional pre-training methods when applied to data-constrained domains and introduce a unified co-regularization approach to address these challenges. Specifically, global and local reconstruc-

tion mechanisms are employed to capture more accurate feature dependencies, while leveraging intra- and inter-modality contrastive learning to derive more generalizable representations for enhanced similarity measurement.

- For fine-tuning, we propose a difference reconstruction mechanism that enables discriminative feature extraction from visually similar medical images under data-limited conditions. Additionally, we introduce a head mixup method to augment data and regularize the model. Comprehensive experiments demonstrate that CMDPI achieves competitive or superior performance to existing models while requiring significantly reduced pre-training data.

This paper is organized as follows: Section 2 reviews related work on medical VQA. Section 3 introduces the overall framework of CMDPI and the key techniques employed in both pre-training and fine-tuning stages. Section 4 presents experimental results and analysis, while Section 5 concludes this paper.

2. Related Work

Early medical VQA models [1, 36] were typically pre-trained on large-scale general datasets such as ImageNet [8] or COCO [28] and then fine-tuned on medical datasets. This transfer learning approach leverages abundant natural image knowledge to alleviate the scarcity of annotated medical data [42]. However, there exist significant domain gaps between natural and medical images in imaging principles, grayscale resolution, tissue complexity, and subtle pathological changes [20, 2, 3]. As a result, directly transferring features from natural images often fails to capture critical medical details, leading to limited performance in medical VQA tasks.

Subsequently, researchers shifted their focus from general-domain data to medical-specific pre-training. MEVF [40] trained an unsupervised convolutional denoising auto-encoder [37] on over 10,000 medical images, combined with model-agnostic meta-learning [10] to initialize the feature extractor. Due to the persistent data limitation in medical VQA, pre-training frameworks began incorporating diverse medical datasets such as report generation [18, 19, 41] and single-modal image datasets [4, 43]. The scale of pre-training data has steadily increased across studies: CMSA [11] used over 10k images, CPRD [29] scaled to approximately 23k unlabeled radiological images, M2I2 [25] utilized 80k samples from ImageCLEF2022, followed by PubMedClip [9], WSRP [14], and VB-MVQA [6] with over 80k image-caption pairs, and MUMC [24] with nearly 400k captions.

In recent years, the emergence of large language models (LLMs) has driven the field toward generative approaches. LLaVA-Med [23] curated 600k biomedical figure-caption pairs, PeFoMed [32] employed over 750k samples, and BiomedCLIP [52] was pre-trained on 15 million image-text pairs. While these generative-based models achieve better parameter reuse and open-ended answer generation compared to traditional classification-based models, they are highly dependent on massive paired datasets and tend to suffer from hallucinations or poor generalization when data is limited.

Although medical-specific pre-training generally provides better domain alignment than general image-text pre-training (e.g., CLIP adaptations), it scales poorly due to the slow accumulation of high-quality medical data caused by privacy, ethical, and expertise constraints. In contrast, general-domain pre-training offers abundant data but suffers from domain gaps.

To address data scarcity more directly, meta-learning and few-shot learning techniques have been explored, such as MAML adaptations [10] and few-shot learner [38]. These methods aim to enable fast adaptation with very few examples. However, they often require diverse meta-datasets or high-quality prompts, which are difficult to obtain in privacy-sensitive medical settings, and may not sufficiently capture fine-grained discriminative patterns across modalities. CMDPI complements these approaches by fundamentally optimizing the training paradigm itself to achieve strong performance under limited data without relying heavily on scaling or post-hoc adaptation.

In summary, although existing methods have progressively increased pre-training data scale to push performance boundaries, they face inherent limitations due to the slow accumulation of medical data. Improving data utilization efficiency under limited-data constraints remains a critical challenge, which this work addresses through paradigm redesign rather than indiscriminate data scaling.

3. Model

This section provides a detailed introduction to the core concepts and key technologies of CMDPI. Firstly, a systematic elaboration of the overall framework of CMDPI is provided. Subsequently, we make full use of the advantage of the framework's compatibility with various pre-training techniques by using multiple pre-training techniques together: global and local reconstruction, as well as intra- and inter-modality contrastive learning. This method facilitates the model of using the idea of co-regularization to enhance its generalizability under limited data conditions. Finally, we discuss a critical fine-tuning technique—differential reconstruction—which facilitates the model's ability to identify discriminative patterns in medical VQA.

The overall framework of CMDPI is designed to effectively integrate multimodal information through co-regularization, thereby achieving more generalized feature

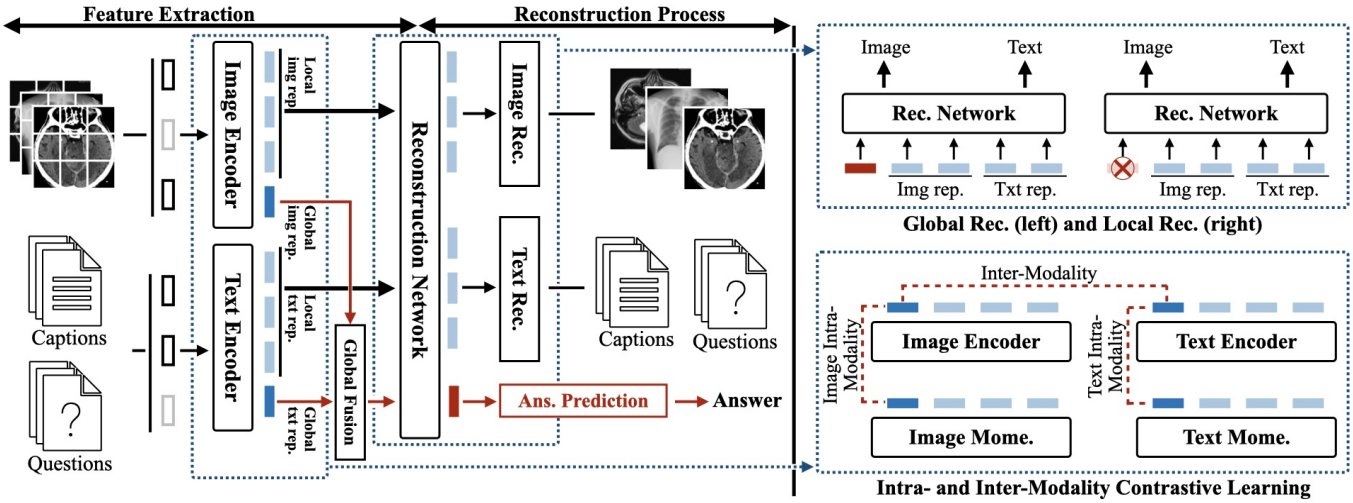


Figure 2: General Structure of the Cross-Modality Discriminative Pattern Identification Model (CMDPI) for Medical VQA.

representations. By introducing co-regularization at different stages, CMDPI can exploit the complementary information among modalities and enhance the performance of downstream tasks.

3.1. Framework

To switch from pre-training tasks to downstream VQA task, existing classification-based models have flexible structures which can be compatible with various pre-training techniques, but they have a newly-added module which cannot benefit from the pre-training process. In contrast, generation-based methods can fully make use of parameters from pre-training by providing a unified framework to process data in both pre-training and fine-tuning stages, but its performance in downstream tasks highly rely on data scale [49].

Considering the low data efficiency of existing methods, our work introduces a **Cross-Modality Discriminative Pattern Identification model (CMDPI)**. As shown in Figure 2, we redesign the model framework so it can incorporate the advantages from both classification-based methods and generation-based methods, which means it is compatible with existing pre-training techniques and also can fully make use of parameters learned from pre-training process. To achieve this, it requires: 1) flexible enough structures to encode global and local features from different modalities to support various pre-training techniques; 2) a unified process that can include both pre-training and fine-tuning stages. Thus, there are three modules in CMDPI: text encoder, image encoder, and reconstruction network. The first two modules are used to extract features from image and text data, while the reconstruction network is used to integrate these features and reconstruct the original data. The encoders extract both global and local features for input data so the model is compatible with various pre-training techniques. And the reconstruction network unifies question answering as a part of the reconstruction task, so the model can fully make use of the parameters obtained from the pre-training stage. All these three modules are constructed based on Transformer.

More specifically, given an image and text input, our method first transform them into features: $\mathrm{Input}_I = [\mathrm{head}_I, I_1, I_2, ..., I_N]$, and $\mathrm{Input}_T = [\mathrm{head}_T, T_1, T_2, ..., T_M]$. Then, the models encode and fuse image and text features as follows:

$$\begin{aligned} \mathbf{h}_I,\ \mathbf{H}_I &= \mathrm{ImageEncoder}\big(\mathrm{Input}_I\big), \\ \mathbf{h}_T,\ \mathbf{H}_T &= \mathrm{TextEncoder}\big(\mathrm{Input}_T\big), \\ \mathbf{h}_F &= \mathrm{GlobalFusion}\big(\mathbf{h}_I, \mathbf{h}_T\big), \end{aligned} \tag{1}$$

where image encoder and text encoder is both based on the Transformer structure [47] in this work, $\mathbf{H}_I \in \mathbb{R}^{N\times d}$ and $\mathbf{H}_T \in \mathbb{R}^{M\times d}$ are the local features of Input_I and Input_T; $\mathbf{h}_I$ and $\mathbf{h}_T$, the corresponding results of head_I and head_T, are global representations, respectively. These two features are then fed into the global fusion module (which is a feed-forward network) to obtain the fused representation $\mathbf{h}_F$. And $\mathbf{h}_F$ is then combined with local features and passed to the reconstruction network to jointly regenerate the original image and text data.

There are two training stages for CMDPI: pre-training and finetuning. The pre-training stage requires the model to learn the joint dependencies between image and text data through pre-training and also obtain aligned representations across different modalities. This process is essential for the model to effectively capture the complex relationships between medical images and their corresponding textual descriptions. For the finetuning, the model needs to effectively

identify discriminative patterns in medical images to answer questions accurately. In the following, we will introduce the key techniques used in these two stages.

3.2. Co-Regularized Pre-training

Existing pre-training techniques are originally designed for general domains with abundant data, and their effectiveness is highly dependent on large-scale training datasets. When applied to data-constrained domains, such as medical VQA, these methods exhibit significant limitations. To address these challenges, we fully make use of the advantage of the framework's compatibility with various pre-training techniques by using the idea of co-regularization. More specifically, our model integrates global-local reconstruction with intra- and inter-modality contrastive learning, so different methods can regularize each other and enhance model generalizability.

3.2.1 Global and Local Reconstruction

Incorporating reconstruction mechanisms during pre-training enhances downstream task performance by enabling models to capture intricate data dependencies. However, under conditions of sparse token distributions, these learned dependencies may exhibit reduced accuracy and reliability. To address this limitation, both global and local reconstruction approaches are incorporated, allowing the dependencies learned at different granularities to mutually regularize and refine each other. These complementary reconstruction methodologies are illustrated in the top right area of Figure 2 and are detailed as follows:

$$\begin{aligned} \mathbf{h}'_I,\ \mathbf{H}'_I &= \mathrm{ImageEncoder}\big(\mathrm{Mask}(\mathrm{Input}_I, \rho)\big), \\ \mathbf{h}'_T,\ \mathbf{H}'_T &= \mathrm{TextEncoder}\big(\mathrm{Mask}(\mathrm{Input}_T, \rho)\big), \\ \mathbf{H}^g_o &= \mathrm{RecNet}\big(\mathbf{h}_F, \mathbf{H}'_I, \mathbf{H}'_T\big), \\ \mathbf{H}^l_o &= \mathrm{RecNet}\big(\tilde{\mathbf{h}}, \mathbf{H}'_I, \mathbf{H}'_T\big), \end{aligned} \tag{2}$$

where ρ is the mask ratio, $\mathbf{H}^g_o$ represents the hidden representations from global reconstruction that obtain reconstructed results based on the fused global vector $\mathbf{h}_F$ (which is derived from the input in Equation 1 without masking), while $\mathbf{H}^l_o$ is calculated through local reconstruction based on a shared trainable vector $\tilde{\mathbf{h}}$. The key distinction between these two reconstruction methods lies in their utilization of global information. Global reconstruction leverages the global information within $\mathbf{h}_F$ to guide the model in capturing dependencies between global and local features, whereas local reconstruction deliberately masks global information by employing a shared vector $\tilde{\mathbf{h}}$, thereby enabling the model to focus exclusively on capturing local feature dependencies.

Based on Transformer [47], the reconstructing process is calculated as follows:

$$\begin{aligned} \mathbf{v}_0 &= [\mathbf{h}, \mathbf{H}_I, \mathbf{H}_T] + pos, \\ \hat{\mathbf{v}}_l &= \mathrm{LN}(\mathrm{MHA}(\mathbf{v}_l) + \mathbf{v}_l), \\ \mathbf{v}_{l+1} &= \mathrm{LN}(\mathrm{FFN}(\hat{\mathbf{v}}_l) + \hat{\mathbf{v}}_l), \end{aligned} \tag{3}$$

where pos is the positional embedding, $\mathrm{MHA}(\cdot)$ denotes the multi-head attention mechanism (implemented as self-attention in our model), $\mathrm{FFN}(\cdot)$ denotes the feed-forward network, where the activation function employed is GELU [15], and $\mathrm{LN}(\cdot)$ is layer normalization [5]. Upon obtaining $\mathbf{H}_o = \mathbf{v}_L$ from the final layer, specialized image, text, and answer reconstruction layers are employed to reconstruct the respective data based on their corresponding feature representations. This process is formulated as follows:

$$\begin{aligned} \mathrm{Output}_I &= \mathrm{FFN}_I(\mathbf{H}_o[:, 1:N]), \\ \mathrm{Output}_T &= \mathrm{FFN}_T(\mathbf{H}_o[:, N+1:N+M]), \\ \mathrm{Ans}' &= \mathrm{FFN}_A(\mathbf{H}_o[:, -1]), \end{aligned} \tag{4}$$

where Output_I, Output_T, and Ans' represent the reconstructed image, text, and answer, respectively. The model is trained to minimize the differences between the reconstructed outputs and the original inputs using Mean Squared Error (MSE) loss for image reconstruction and cross-entropy loss for text and answer reconstruction. The overall reconstruction loss is defined as follows:

$$\begin{aligned} \mathbf{L}_{img} &= \mathrm{MSE}(\mathrm{Output}_I, \mathrm{Input}_I), \\ \mathbf{L}_{txt} &= \mathrm{CrossEntropy}(\mathrm{Output}_T, \mathrm{Input}_T), \\ \mathbf{L}_{ans} &= \mathrm{CrossEntropy}(\mathrm{Ans}', \mathrm{Ans}), \\ \mathbf{L}_{rec} &= \mathbf{L}_{img} + \mathbf{L}_{txt} + \mathbf{L}_{ans}, \end{aligned} \tag{5}$$

where Ans denotes the target answer. By integrating both global and local reconstruction mechanisms, the model can effectively capture dependencies at multiple granularities, thereby enhancing its ability to learn robust and generalizable feature representations even under conditions of limited data. It should be noted that the QA data used in pre-training is generated automatically. Currently, high quality medical VQA data are still limited and we only used it in the finetuning stage.

3.2.2 Intra- and Inter-Modality Contrastive Learning

Contrastive learning, which enables model to measure data similarities, can help further regularize the hidden representations from the image and text encoders. However, the similarity measurements learned in limited data are not generalizable enough to support downstream tasks. We thus adopt both Intra-modality and Inter-modality contrastive learning, so the similarity within same modalities can regularize

the representations across different modalities. The intra-modality is implemented based on the momentum contrastive learning [13], which maintains a momentum encoder for each modality to provide more consistent representations as learning targets.

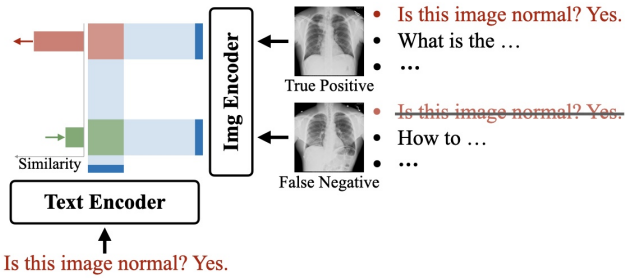


Figure 3: Analysis of VQA Data in Contrastive Learning.

CMDPI is thus able to leverage the complementary information present in both modalities, leading to more robust and generalizable feature representations. This process is shown in the bottom right of Figure 2. Given the global representations from the momentum image encoder $\hat{\mathbf{h}}_I$ and momentum text encoder $\hat{\mathbf{h}}_T$, the contrastive learning used in our work is:

$$\begin{aligned}\mathbf{C}_{inter} &= \text{Contra}\big(\mathbf{h}_I, \mathbf{h}_T\big) + \text{Contra}\big(\mathbf{h}_T, \mathbf{h}_I\big),\\ \mathbf{C}_{intra} &= \text{Contra}\big(\mathbf{h}_I, \hat{\mathbf{h}}_I\big) + \text{Contra}\big(\mathbf{h}_T, \hat{\mathbf{h}}_T\big),\\ \mathbf{C}_{total} &= \alpha \cdot \mathbf{C}_{inter} + \alpha' \cdot \mathbf{C}_{intra},\end{aligned} \tag{6}$$

where α and α' are two scaling factors to adjust the importance of $\mathbf{C}_{inter}$ and $\mathbf{C}_{intra}$, which are set to be 0.5 and 1, respectively. This kind of design allows the model to jointly learn data similarities of both same and different modalities in a same representation space. As a result, CMDPI becomes better equipped to capture complex multimodal relationships, ultimately enhancing its ability to perform the downstream task with improved accuracy and reliability. The contrastive learning used is as follows:

$$\text{Contra}\big(a, b\big) = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(sim(a^i, b^i)/\tau)}{\sum_{j=1}^{N} exp(sim(a^i, b^j)/\tau)}, \tag{7}$$

where N denotes the number of image-text pairs, and $\text{sim}(a^i, b^i) = \frac{{a^i}^T b^i}{\|a^i\|\|b^i\|}$ represents the precomputed similarity. a and b are general representations referring to $\mathbf{h}_I$, $\mathbf{h}_T$, $\hat{\mathbf{h}}_I$, $\hat{\mathbf{h}}_T$ in Equation 6.

3.2.3 Compatibility of VQA Data in Contrastive Learning

There are two kinds image-text medical data: medical image captions and medical VQA data. One straight-forward way is to incorporate these two kinds of data in both reconstruction and contrastive learning during the pre-training stage. However, it is observed that there is a performance decrease after incorporating the VQA data into contrastive learning. The main reason is from the incomplete information of VQA data.

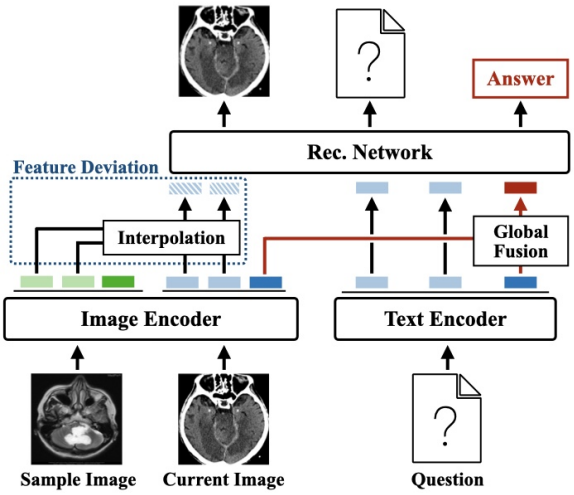


Figure 4: The Process of Difference Reconstruction.

In contrastive learning, the model will increase its similarity to positive data, while decrease the similarity to negative data. Different with image captions providing complete description of an image, each VQA sample only contains one aspect of an image. Although each image may have multiple VQA samples, they do not have a guarantee to describe an image completely. It is thus easy for VQA data to have false negative samples during contrastive learning. An example is given in Figure 3. Given two images without abnormalities, the first image contains a VQA sample describes it, while the second image does not have such data. For the question describing the image normality, the second image will be a false negative samples, and its similarity to the given data will be decreased in the contrastive process. Such noises led by the fake negative samples will lead to misalignment between image and text data and finally influence model performance in downstream tasks.

Therefore, for the pre-training of CMDPI, our paper adopts both reconstruction and contrastive learning for caption data, and only use reconstruction for VQA data.

3.3. Fine-tuning

After pre-training, fine-tuning is necessary to fit the model into downstream medical VQA tasks. Our framework has a specific head to predict answer during pre-training, and it can be used directly in the fine-tuning stage.

Medical images frequently exhibit high visual similarities, while the key to accurate question answering lies in

capturing distinctive features unique to individual cases. Given the extreme challenges of learning such cross-modal patterns under limited data conditions, our paper adopts a **difference reconstruction** mechanism to capture discriminative differences among similar images. The core principle involves prompting the model to reconstruct target images from similar reference images, thereby enabling it to learn the distinguishing characteristics between the current image and similar counterparts. These differences consistently represent the most discriminative and diagnostically important features. More specifically, given an input image, it is first encoded into a global feature representation $\mathbf{h}_I$ and local feature representations $\mathbf{H}_I$ using Equation 1. Subsequently, we sample another similar image and extract its local features $\hat{\mathbf{H}}_I$, followed by the application of a feature deviation operation. Given the current image Input_I and another similar image $\hat{\mathrm{Input}}_I$, this process is demonstrated in Figure 4 and calculated as follows:

$$\begin{aligned}
&\mathbf{h}_I,\ \mathbf{H}_I = \mathrm{ImageEncoder}\big(\mathrm{Input}_I\big),\\
&\hat{\mathbf{h}}_I,\ \hat{\mathbf{H}}_I = \mathrm{ImageEncoder}\big(\hat{\mathrm{Input}}_I\big),\\
&\tilde{\mathbf{H}}_I = \epsilon \cdot \mathbf{H}_I + (1-\epsilon)\cdot \hat{\mathbf{H}}_I,\\
&\mathbf{H}_o^d = \mathrm{RecNet}\big(\mathbf{h}_F, \tilde{\mathbf{H}}_I, \mathbf{H}_T\big),
\end{aligned} \tag{8}$$

where $\mathbf{H}_T$ is the local text representations and $\mathbf{h}_F$ is the fused global representation from Equation 1. ϵ is randomly sampled from $[0,1]$ and $\tilde{\mathbf{H}}_I$ is the deviated local features obtained by combining local features from two similar images. Then, the reconstruction network will follow Equation 4 to finish the reconstruction based on $\mathbf{H}_o^d$, which is:

$$\begin{aligned}
&\mathrm{Output}_I^d = \mathrm{FFN}_I(\mathbf{H}_o^d[:, 1:N]),\\
&\mathrm{Output}_T^d = \mathrm{FFN}_T(\mathbf{H}_o^d[:, N+1:N+M]).
\end{aligned} \tag{9}$$

And the training objective of the difference reconstruction is:

$$\begin{aligned}
\mathbf{L}_{img}^d &= \mathrm{MSE}(\mathrm{Output}_I^d, \mathrm{Input}_I),\\
\mathbf{L}_{txt}^d &= \mathrm{CrossEntropy}(\mathrm{Output}_T^d, \mathrm{Input}_T),\\
\mathbf{L}_{diff}^d &= \mathbf{L}_{img}^d + \mathbf{L}_{txt}^d.
\end{aligned} \tag{10}$$

This operation has two advantages. First, it prompts the model to compare two similar images and learn their differences, so that the unique features of each image can be captured in this process. Then, the random factors inherent in this process can prevent the model from relying too heavily on the original features, thereby reducing the risk of overfitting. This forces explicit learning of fine-grained subtle differences in high-similarity medical images. This method integrates with co-regularized pre-training, and better addresses the core challenge of data-scarce medical VQA.

Nevertheless, the extremely limited data during fine-tuning require further data augmentation and regularization to prevent overfitting. Based on the idea of Mixup [50], a head mixup method for CMDPI is introduced. Given the local features $\mathbf{H}_o^d$ from difference reconstruction, we first obtain its downstream head $\mathbf{h}_D = \mathbf{H}_o^d[:, -1]$, then we pick another VQA sample and obtain its downstream head $\dot{\mathbf{h}}_D = \dot{\mathbf{H}}_o[:, -1]$. We then mix these two heads and their corresponding answers Ans and $\dot{\mathrm{Ans}}$ as follows:

$$\begin{aligned}
\mathbf{h}_{mix} &= \lambda \cdot \mathbf{h}_D + (1-\lambda)\cdot \dot{\mathbf{h}}_D,\\
\mathrm{Ans}_{mix} &= \lambda \cdot \mathrm{Ans} + (1-\lambda)\cdot \dot{\mathrm{Ans}},
\end{aligned} \tag{11}$$

where λ is randomly sampled from $[0,1]$. The model is trained to minimize $\mathrm{Ans}'_{mix} = \mathrm{FFN}_A(\mathbf{h}_{mix})$ and the mixed label Ans_{mix} via cross-entropy loss, which is:

$$\mathbf{L}_{mix} = \mathrm{CrossEntropy}(\mathrm{Ans}'_{mix}, \mathrm{Ans}_{mix}). \tag{12}$$

Combining the loss functions from difference reconstruction and head mixup, the overall loss function during fine-tuning is:

$$\mathbf{L}_{ft} = \mathbf{L}_{mix} + \mathbf{L}_{diff}^d. \tag{13}$$

Head mixup can learn more generalizable features while difference reconstruction sharpens feature discriminability. It should be noted that the hidden representations generated through feature deviation also contribute to answer prediction. It enables the model to learn discriminative image patterns for question answering. However, to prevent excessive noise amplification, we restrict feature deviation application exclusively to the current sample, excluding the sampled similar image.

4. Experiment

This section begins by presenting the datasets for pre-training and fine-tuning. Next, the experimental setup is described, including the compared models and hyper-parameters. Experimental results are then demonstrated, followed by a discussion that further analyzes the CMDPI.

4.1. Dataset

The ROCO dataset [41] comprises a multimodal collection featuring a variety of medical imaging modalities (such as X-rays, CT, MRI.), encompassing over 81k medical radiographic images accompanied by their corresponding textual descriptions. PMC-VQA [53] is a large-scale medical VQA dataset, which contains 149k images with 227k QA pairs covering various modalities or diseases. The QA pairs are automatically generated using ChatGPT. And these images encompass various medical imaging modalities, fulfilling the requirement for dataset diversity. These two datasets are utilized in our pre-training process, while we only uses the training split in ROCO: 65k data and a small part of

Table 1: Model Performance on VQA-RAD Dataset. Models with * are based on LLMs.

Model	Open	Closed	Overall
Pretrain Data Less than 120k			
MEVF [40]	43.0%	76.5%	63.2%
VQA-MIX [12]	63.1%	82.4%	74.7%
MLPs-Base [34]	53.1%	81.3%	70.2%
VG-CALF [21]	67.0%	85.5%	76.1%
ACMA-MAM [26]	63.6%	84.4%	76.1%
CPCR [31]	60.5%	80.4%	72.5%
WSRP [14]	67.6%	82.7%	76.7%
VQA-Adapter [33]	66.1%	82.3%	75.8%
MKGF [48]	61.7%	81.7%	71.7%
VB-MVQA [6]	61.3%	78.1%	71.3%
MITER [45]	59.4%	80.5%	72.1%
CMDPI (Our model)	74.3%	85.7%	**78.7%**
Pretrain Data More than 200k			
M^3AE [7]	67.2%	83.5%	77.0%
BiomedCLIP [52]	67.0%	76.5%	72.7%
LLaVA-Med* [23]	61.5%	84.2%	75.2%
UniMed-LVLM*[51]	26.2%	75.4%	55.8%
PeFoMed* [32]	62.6%	87.1%	**77.4%**

Table 2: Model Performance on SLAKE Dataset. Models with * are based on LLMs.

Model	Open	Closed	Overall
Pretrain Data Less than 120k			
MEVF [40]	76.1%	81.7%	78.3%
VQA-MIX [12]	77.2%	80.8%	78.6%
MLPs-Base [34]	79.9%	88.5%	82.9%
VG-CALF [21]	81.4%	83.8%	83.3%
ACMA-MAM [26]	80.8%	86.7%	83.1%
CPCR [31]	80.5%	84.1%	81.9%
WSRP [14]	76.7%	83.9%	79.5%
VQA-Adapter [33]	79.2%	83.7%	81.0%
MKGF [48]	65.8%	82.0%	74.3%
VB-MVQA [6]	77.0%	80.5%	78.7%
MITER [45]	79.2%	84.4%	81.2%
CMDPI (Our model)	80.6%	88.0%	**83.5%**
Pretrain Data More than 200k			
M^3AE [7]	80.3%	87.8%	83.2%
BiomedCLIP [52]	84.3%	88.9%	**86.1%**
LLaVA-Med* [23]	83.1%	85.3%	84.0%
UniMed-LVLM*[51]	84.0%	84.4%	84.2%
PeFoMed* [32]	77.8%	88.7%	82.1%

PMC-VQA: 49k, which is 114k in total to pre-train our model.

To evaluate our proposed method, two public-available Med-VQA benchmark datasets are selected. The VQA-RAD dataset [22] is a medical VQA dataset manually curated in radiology, comprising 315 radiology images and 3,515 visual questions, question-answer pairs are categorized into open-ended tasks and closed-ended tasks. 42% of which were open-ended and the rest were closed-ended. The SLAKE dataset [30] is a comprehensive medical VQA dataset which comprises 642 medical images and 14,028 question-answer pairs. 60% of the questions were open-ended, while the remaining were closed-ended.

4.2. Experimental setup

Our paper compares the performance of CMDPI with existing mainstream models in medical VQA, including MEVF [40], VQA-MIX [12], MLPs-Base [34], VG-CALF [21], ACMA-MAM [26], CPCR [31], WSRP [14], VQA-Adapter [33], MKGF [48], VB-MVQA [6], MITER [45] , M^3AE [7], BiomedCLIP [52], LLaVA-Med [23], UniMed-LVLM [51], and PeFoMed [32]. These models encompass both classification-based and large language model-based approaches for medical VQA, each exhibiting unique characteristics.

For our proposal model, CMDPI, its image encoders and text encoders are initialized with CLIP [44]. The reconstruction network adopts Transformer [47] as backbone. The layer number is set to be 4 with 512 hidden units. During pre-training, AdamW is used as optimizer. The maximum number of training epochs is 100. During fine-tuning, the learning rate is set to 7.5e-6, with a maximum of 300 epochs. The batch size is fixed at 32.

During difference reconstruction, we identify similar image pairs using our pre-trained model (prior to fine-tuning), leveraging its generalized capability for similarity measurement. For each image, we extract a global representation. Pairs exhibiting a cosine similarity exceeding 0.75 are treated as similar. When employing the head mixup technique, the sampled data maintain consistent organ type and question type with the original samples. The model comprises approximately 194 million parameters. Pre-training required 20 hours, while fine-tuning took 14 hours on a single NVIDIA RTX 3090 GPU.

4.3. Experimental Results

The results on the VQA-RAD and SLAKE datasets are shown in Table 1 and 2, respectively. Generally, there is no clear gap between existing classification-based models and LLM-based models. Although LLM-based models have large volume of data pre-training on huge general-domain dataset, they cannot benefit from this process. It demonstrates the high differences between image data and general domain data.

Existing models can be broadly categorized into two groups based on the volume of pre-training data: pre-

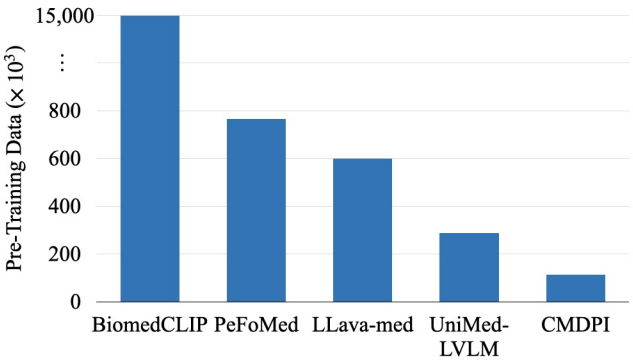


Figure 5: Number of Pre-training Data.

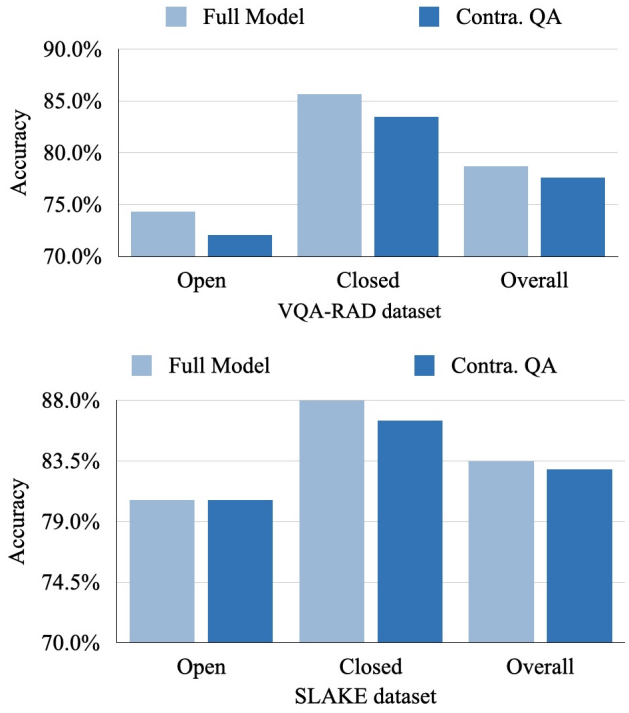


Figure 6: Ablation Study of Contrastive Learning Data.

training data less than 120k and pre-training data more than 200k. Within the first category (models with limited pre-training data), CMDPI achieves satisfactory accuracy on both the VQA-RAD and SLAKE datasets, surpassing existing benchmarks including powerful LLM-based models. This enhancement can be attributed to three key innovations: (1) the integration of co-regularization during pre-training, which strengthens the model's ability to learn generalizable cross-modality alignment features; (2) the high utilization of pre-trained parameters, which increases the model's capacity to capture complex patterns; and (3) the incorporation of difference reconstruction and mixup techniques during fine-tuning, which enables efficient optimization with limited data for downstream tasks. The proposed approach not only elevates model accuracy but also reduces reliance on large-scale datasets.

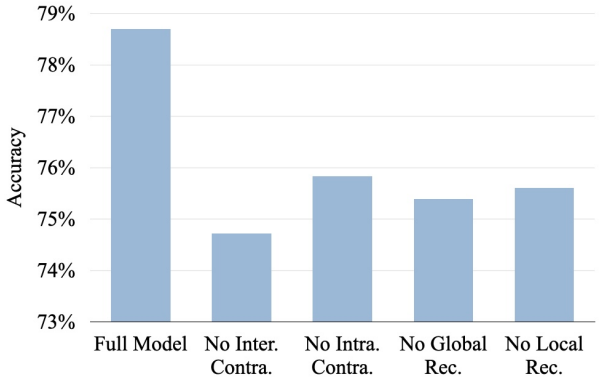


Figure 7: Ablation Study of Pre-training Techniques.

When benchmarked against models pre-trained on over 200,000 data samples, CMDPI achieves the best accuracy of 78.7% on the VQA-RAD dataset, outperforming existing approaches. On the SLAKE dataset, although BiomedCLIP achieved the best results, in terms of the amount of pre-training data, as shown in Figure 5, it leverages 15M pre-training data samples—over 13 times the volume required by CMDPI. Furthermore, BiomedCLIP achieves 6% lower accuracy than our model on VQA-RAD. Similarly, while UniMed-LVLM performs well on SLAKE, its accuracy on VQA-RAD drops to 55.8%, and the amount of pre-training data is higher than that of CMDPI. This disparity underscores BiomedCLIP and UniMed-LVLM's limited generalization capabilities. In contrast, CMDPI requires substantially less data while demonstrating superior generalization performance compared to models that depend on significantly larger datasets. The superior performance of CMDPI relative to existing classification-based and LLM-based methods demonstrates that it represents a promising approach for enhancing performance through optimized model architectures and training strategies rather than through continuous data scaling.

4.4. Discussion

In addition, ablation studies are conducted to verify the effectiveness of our design in both pre-training and fine-tuning.

4.4.1 Effectiveness of Pre-training Techniques

Due to the inherent incompleteness of information in VQA datasets, they are not well-suited for contrastive learning. VQA data often lack comprehensive and detailed annotations, and the available question-answer pairs may not fully capture the rich semantic relationships present in the corre-

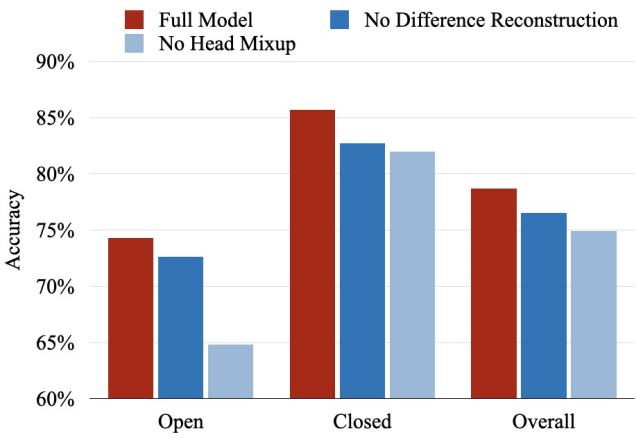


Figure 8: Ablation Study of Fine-tuning Techniques.

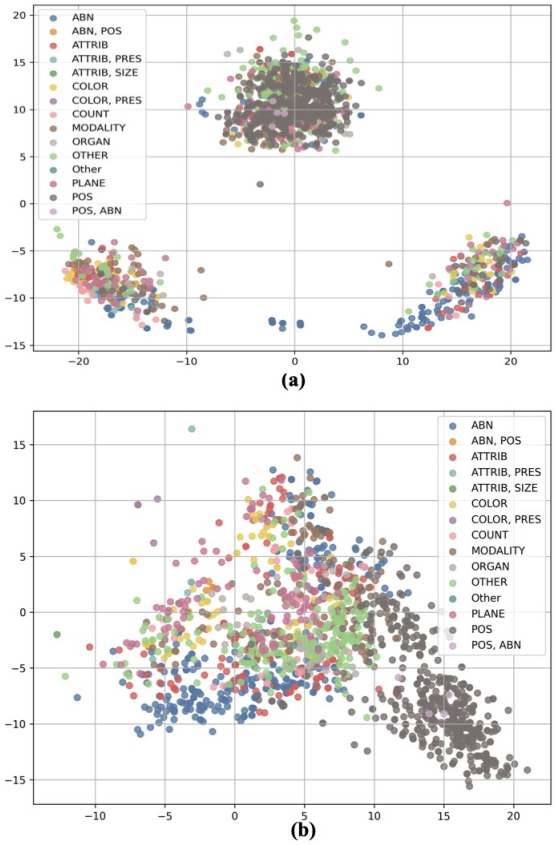


Figure 9: Feature Visualization of (a) CMDPI without head mixup; and (b) CMDPI with head mixup.

sponding images and texts. As a result, applying contrastive learning directly on VQA data can lead to suboptimal alignment between modalities and insufficiently discriminative feature representations. The experimental results are shown in Figure 6. The dark blue bar (labeled as Contra.QA) corresponds to the model that incorporates VQA data in

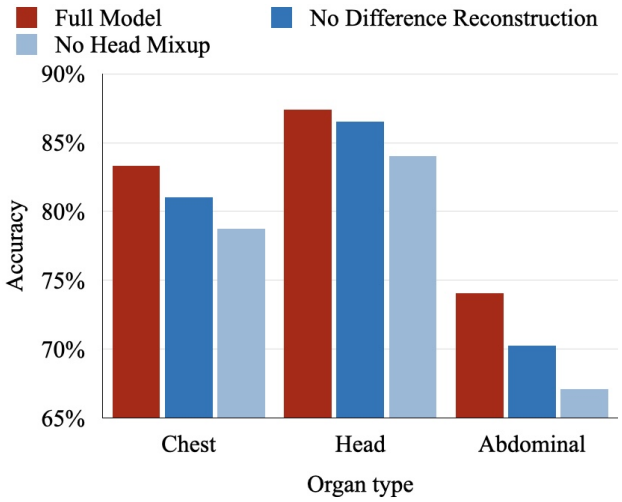


Figure 10: Ablation Fine-Tuning Performance on Different Organ Types.

contrastive learning, and its performance shows a performance drop. This observation is consistent with our analysis, which suggests that the limited and incomplete information in VQA data hinders the effectiveness of contrastive learning strategies, ultimately affecting the model's ability to generalize to downstream tasks.

Furthermore, we integrate the idea of co-regularization, which are global and local reconstruction, and inter- and intra-modality contrastive learning in pre-training process. It can be observed in Figure 7 that disabling any of them results in a clear decrease in performance. Under the setting of limited pre-training data, the original pre-training techniques cannot obtain good performance as expected. It is important to apply regularization for each of them. Our core idea, co-regularization contributes to the performance of CMDPI. Comparing to these four techniques, inter-modality contrastive learning plays a more important role, since the performance decrease more than the other three techniques. It shows the significance of the alignment between different modalities.

4.4.2 Effectiveness of Fine-tuning Techniques

The effectiveness of difference reconstruction and head mixup is illustrated in Figure 8. Difference reconstruction can contribute to both the model performance in open questions and closed questions, while head mixup contributes more to the performance in open questions. Comparing with closed questions, open questions have more various answers. In this time, head mixup can be regarded as a regularization term to ensure a rough linear relations between selected answers. To demonstrate this analysis, our paper further conducts a visualization of the features from

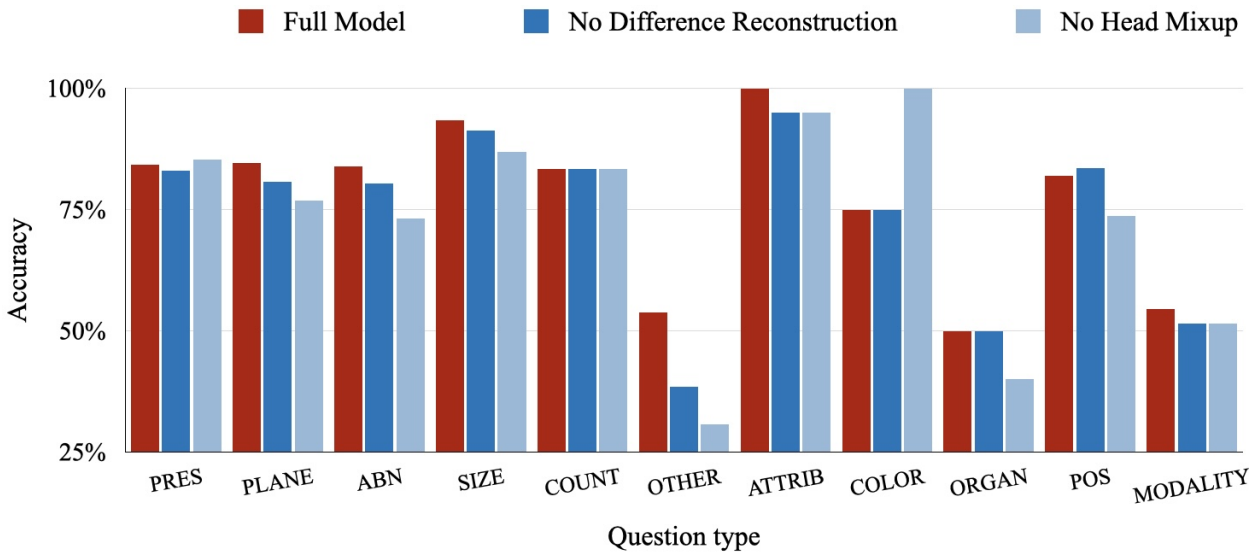


Figure 11: Ablation Fine-Tuning Performance on Different Question Types.

the model with and without head mixup. The results are shown in Figure 9. In Figure 9 (a), data with different question types are in different colors. The model without head mixup has large gaps between different question types. Features corresponding to different questions types are mixed and divided into three clusters. There are no clear patterns in each cluster. However, in Figure 9 (b), the model with head mixup has a more compact distribution and features corresponding to different questions types have been distributed in certain areas. These results verify of our analyses to head mixup that head mixup can be regarded as a regularization technique, and the model with head mixup can learn more generalizable features.

To further explore the effects of difference reconstruction and head mixup in various specific scenarios, we conducted more detailed ablation analyses by categorizing the performance metrics according to question type and organ type, as shown in Figure 10 and Figure 11. As illustrated in Figure 10, the overall trend is largely consistent with the results in Figure 8, with our model achieving the best performance across all metrics. And the difference reconstruction technique performs better than head mixup. It is worth noting that the performance on abdominal-related questions is lower than that on head and chest-related questions, which may be attributed to the complexity of abdominal anatomy.

As shown in Figure 11, the CMDPI model demonstrates superior performance across most question types, while the variant models that employ only difference reconstruction or head mixup achieve higher efficiency in only a few categories, such as color and position. This phenomenon may be attributed to the fact that these question types mainly focus on factual issues, requiring direct extraction of basic visual attributes and involving less complex cross-modal reasoning. In contrast, for higher-level reasoning questions such as abnormality, our CMDPI model stands out. This is primarily because such questions rely on cross-modal semantic understanding and complex clinical reasoning. The synergy between difference reconstruction and head mixup enables the model to enhance fine feature discrimination and answer generalization under limited data conditions, resulting in performance gains that surpass those of single techniques.

More experimental results are provided in the Appendix.

4.5. Case Study

In addition, we present a case study to explore the effectiveness of difference reconstruction and head mixup, as illustrated in Figure 12. The first case is a closed-ended question, for which the target answer should be ”No.” However, only our full model is able to provide the correct answer. The model without difference reconstruction fails to capture key information and thus cannot produce the correct answer. The model without head mixup learns less diverse features and provides the wrong answer. The second case involves an open-ended answer type, relating to a question about size. Both the model without difference reconstruction and the model without head mixup fail to correctly identify the answer type, mistakenly treating it as a closed-ended task and outputting an incorrect answer, whereas CMDPI obtains the correct answer. The randomness introduced by difference reconstruction prevents the model from over-relying on latent features, while head mixup alleviates overfitting. The combination of these two mechanisms not only helps the model capture discriminative image features but also fa-

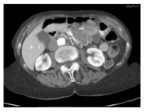
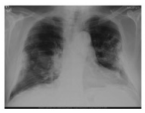
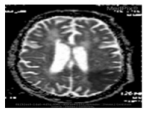
ID	Image	Data			
1		**Question:** Is the surrounding phlegmon normal?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		No ✓	Yes ✗	Yes ✗	No
2		**Question:** How large is the mass?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		5mm ✓	Yes ✗	Yes ✗	5mm
3		**Question:** Where is the diffuse pleural thickening?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		Right lung ✓	Right lung ✓	Yes ✗	Right lung
4		**Question:** Does this look like a healthy liver?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		Yes ✓	Yes ✓	No ✗	Yes
5		**Question:** Is there flattening of the left hemidiaphragm?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		Yes ✓	No ✗	Yes ✓	Yes
6		**Question:** What side of the brain is a lesion on?			
		Full Model	**w/o diff. rec.**	**w/o head mixup**	**Ground Truth**
		Left ✓	Right ✗	Left ✓	Left

Figure 12: Case Study of Fine-tuning Techniques.

cilitates more accurate extraction of textual representations.

In the third and fourth cases, only the model without the head mixup produces incorrect answers. Head mixup plays a crucial role in enhancing the model's cross-modal feature learning capability. By promoting interaction and fusion among different attention heads, the head mixup mechanism enables the model to better capture the complex relationships between images and text. In the fifth and sixth cases, only the model without difference reconstruction produces incorrect answers. Difference reconstruction effectively captures subtle differences between input samples, thereby improving the model's ability to distinguish between similar samples. By modeling minor variations among inputs, the difference reconstruction mechanism helps the model more precisely understand and represent the detailed associations between images and text.

5. Conclusion

In this work, we begin our analysis by examining prevailing trends in existing medical Visual Question Answering (VQA) research. It is observed that current methodologies emulate the large-scale data pre-training paradigms dominant in general-domain VQA systems. However, the rate of data accumulation in the medical domain lags significantly behind that of general-domain repositories. This discrepancy suggests that heavy reliance on large-scale data pre-training approaches will quickly encounter bottlenecks in development.

We propose CMDPI, a data-efficient model for medical visual question answering (VQA). Our framework combines the strengths of both classification-based and generation-based models, ensuring compatibility with existing pre-training techniques and enabling effective utilization of pre-trained parameters. Through our analysis, we find that general-domain pre-training methods often exhibit performance biases under data-constrained conditions. To address this, CMDPI leverages its flexible architecture and introduces a co-regularization strategy that integrates global and local image reconstruction with both inter- and intra-modality contrastive learning. During fine-tuning, to mitigate the high similarity characteristic of medical images, we design a difference reconstruction module that enhances the model's sensitivity to unique image features. Additionally, a head mixup technique is employed to further improve performance. Experimental results demonstrate that CMDPI achieves performance comparable to or better than existing classification-based and LLM-based models, while requiring substantially less pre-training data.

Acknowledgement

The work was supported by Key Research and Development Program of Heilongjiang Province(2024ZXDXA09), the National Natural Science Foundation of China under (Grant No.62072135), the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation, the National Science Foundation of China under the Special Project Scheme (project no. T254101295).

References

[1] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman. Nlm at imageclef 2018 visual question answering in the medical domain. In *CLEF (working notes)*, pages 1–10, 2018. 3

[2] B. Afrae, D. Yousra, A. Imane, B. A. Mohamed, and A. B. Abdelhakim. A new visual question answering system for medical images characterization. In *Proceedings of the 4th International Conference on Smart City Applications*, pages 1–7, 2019. 3

[3] I. Allaouzi, M. B. Ahmed, and B. Benamrou. An encoder-decoder model for visual question answering in the medical domain. In *CLEF (working notes)*, 2019. 3

[4] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 3

[5] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[6] L. Cai, H. Fang, and Z. Li. Pre-trained multilevel fuse network based on vision-conditioned reasoning and bilinear attentions for medical image visual question answering: L. cai et al. *The Journal of Supercomputing*, 79(12):13696–13723, 2023. 3, 8

[7] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T. Chang. Mapping medical image-text to a joint space via masked modeling. *Medical Image Anal.*, 91:103018, 2024. 8

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[9] S. Eslami, C. Meinel, and G. De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. 3

[10] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3

[11] H. Gong, G. Chen, S. Liu, Y. Yu, and G. Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *Proceedings of the 2021 international conference on multimedia retrieval*, pages 456–460, 2021. 3

[12] H. Gong, G. Chen, M. Mao, Z. Li, and G. Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11):3332–3343, 2022. 8

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. 6

[14] S. He, H. Pan, K. Zhang, C. Gong, and Z. Li. A weak supervision-based robust pretraining method for medical visual question answering. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1957–1960. IEEE, 2023. 1, 3, 8

[15] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[16] J. Huang, Y. Chen, Y. Li, Z. Yang, X. Gong, F. L. Wang, X. Xu, and W. Liu. Medical knowledge-based network for patient-oriented visual question answering. *Information Processing & Management*, 60(2):103241, 2023. 1

[17] X. Huang and H. Gong. A dual-attention learning network with word and sentence embedding for medical visual question answering. *IEEE Transactions on Medical Imaging*, 43(2):832–845, 2023. 1

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 3

[19] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3

[20] E. A. Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010. 3

[21] A. Lameesa, C. Silpasuwanchai, and M. S. B. Alam. Vg-calf: A vision-guided cross-attention and late-fusion network for radiology images in medical visual question answering. *Neurocomputing*, 613:128730, 2025. 8

[22] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 8

[23] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 3, 8

[24] P. Li, G. Liu, J. He, Z. Zhao, and S. Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023. 3

[25] P. Li, G. Liu, L. Tan, J. Liao, and S. Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 3

[26] Y. Li, Q. Yang, F. L. Wang, L.-K. Lee, Y. Qu, and T. Hao. Asymmetric cross-modal attention network with multimodal augmented mixup for medical visual question answering. *Artificial Intelligence in Medicine*, 144:102667, 2023. 8

[27] X. Liang, Y. Wang, D. Wang, Z. Jiao, H. Zhong, M. Yang, and Q. Wang. Leveraging coarse-to-fine grained representations in contrastive learning for differential medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–425. Springer, 2024. 1

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[29] B. Liu, L.-M. Zhan, and X.-M. Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International conference on medical image computing and computer-assisted intervention*, pages 210–220. Springer, 2021. 3

[30] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 8

[31] B. Liu, L.-M. Zhan, L. Xu, and X.-M. Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022. 8

[32] G. Liu, J. He, P. Li, G. He, Z. Chen, and S. Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv preprint arXiv:2401.02797*, 2024. 1, 3, 8

[33] J. Liu, T. Hu, Y. Zhang, Y. Feng, J. Hao, J. Lv, and Z. Liu. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2816–2826, 2023. 8

[34] L. Liu and X. Su. How well apply multimodal mixup and simple mlps backbone to medical visual question answering? In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2648–2655. IEEE, 2022. 8

[35] Y. Liu, B. Chen, S. Wang, G. Lu, and Z. Zhang. Deep fuzzy multiteacher distillation network for medical visual question answering. *IEEE Transactions on Fuzzy Systems*, 32(10):5413–5427, 2024. 1

[36] A. Lubna, S. Kalady, and A. Lijiya. Mobvqa: a modality based medical image visual question answering system. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 727–732. IEEE, 2019. 3

[37] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer, 2011. 3

[38] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3

[39] A. Mudgal, U. Kush, A. Kumar, and A. Jafari. Multimodal fusion: advancing medical visual question-answering. *Neural Computing and Applications*, 36(33):20949–20962, 2024. 1

[40] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran. Overcoming data limitation in medical visual question answering. In *International conference on medical image computing and computer-assisted intervention*, pages 522–530. Springer, 2019. 3, 8

[41] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, proceedings*, pages 180–189. Springer International Publishing, 2018. 3, 7

[42] Y. Peng, F. Liu, and M. P. Rosen. Umass at imageclef medical visual question answering (med-vqa) 2018 task. In *CLEF (working notes)*, pages 1–9, 2018. 3

[43] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. 3

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 8

[45] C. Shu, Y. Zhu, X. Tang, J. Xiao, Y. Chen, X. Li, Q. Zhang, and Z. Lu. Miter: Medical image–text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications*, 238:121526, 2024. 8

[46] T. Van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. Snoek, and M. Worring. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–736. Springer, 2023. 1

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5, 8

[48] Y. Wu, Y. Lu, Y. Zhou, Y. Ding, J. Liu, and T. Ruan. Mkgf: A multi-modal knowledge graph based rag framework to enhance lvlms for medical visual question answering. *Neurocomputing*, 635:129999, 2025. 8

[49] B. Zhang, Z. Liu, C. Cherry, and O. Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024. 1, 4

[50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 7

[51] H. Zhang, M. Zeng, J. Ding, Y. Liang, R. Zheng, Z. Qu, M. Li, and S. Kan. Aligning multimodal biomedical images and language via one large vision-language model. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2874–2879. IEEE, 2024. 8

[52] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 3, 8

[53] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 7

Appendix

Additional Experimental Results

Hyperparameters in Pre-training Loss

During the pretraining process, we employ two scaling factors, α and α', to modulate the relative importance of inter-modality contrastive learning and intra-modality contrastive learning interactions in Equation (6), respectively. To investigate the impact of these scaling factors on model performance, we conduct experiments using different values of α and α'. The results are illustrated in Figure 13. The combination of $\alpha = 0.5$ and $\alpha' = 1$ consistently achieves the best performance in the open-ended, closed-ended, and overall tasks. This finding suggests that, slightly down-weighting cross-modal interactions ($\alpha = 0.5$) while maintaining stronger intra-modal interactions ($\alpha' = 1$) is more beneficial for learning rich, complementary, and well-aligned multimodal representations.

We further analyze the impact of different loss weight for each component of the reconstruction loss on model performance during the pre-training stage. Specifically, we adjusted the weights of the image reconstruction loss (L_{img}), text reconstruction loss (L_{txt}), and answer reconstruction loss (L_{ans}), and shown the experimental results under various weight combinations in Table 3. For each loss component, we test settings with a weight of 0.5 (reducing its influence) and 1.0 (default balanced setting). The results show that the model achieves the best performance when all loss components are set to 1.0. In contrast, imbalanced weighting leads to a decline in model performance, as it introduces bias and affects the model's generalization ability. Overall, a balanced weighting of each reconstruction loss component is crucial for improving the overall performance of the model.

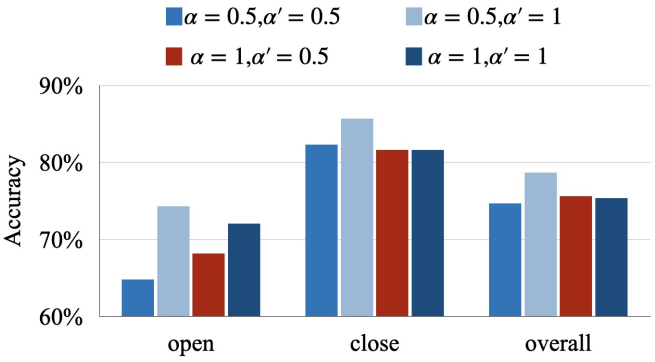


Figure 13: Different Values of Scaling Factor in Contrastive Learning.

Table 3: Ablation Study of Reconstruction Loss Weights in Pretraining on the VQA-RAD Dataset

Loss Weight			Performance		
L_{img}	L_{txt}	L_{ans}	open	close	overall
0.5	0.5	0.5	73.2%	82.4%	76.7%
0.5	0.5	1.0	70.4%	82.4%	76.3%
0.5	1.0	1.0	71.5%	82.7%	76.3%
0.5	1.0	0.5	72.1%	83.5%	76.5%
1.0	0.5	0.5	72.6%	81.3%	76.1%
1.0	0.5	1.0	73.2%	80.5%	75.8%
1.0	1.0	0.5	71.0%	82.0%	76.5%
1.0	1.0	1.0	**74.3%**	**85.7%**	**78.7%**

Table 4: Ablation Study of Finetuning Loss Weights on the VQA-RAD Dataset

Loss Weight		Performance		
L_{mix}	L^d_{diff}	open	close	overall
0.5	0.5	73.2%	82.4%	76.7%
0.5	1.0	70.4%	82.4%	76.3%
1.0	0.5	71.5%	82.7%	76.3%
1.0	1.0	**74.3%**	**85.7%**	**78.7%**

Hyperparameters in Fine-tuning Loss

During the finetuning stage, to evaluate the impact of the weight configuration between the head mixup loss (L_{mix}) and the difference reconstruction loss (L^d_{diff}) on model performance, we conducted an ablation study on the loss weights during the fine-tuning stage of the CMDPI model, as shown in Table 4. The experiment compared four different weight combinations, and the results indicate that the model achieves the best performance when both losses are set to a weight of 1.0. This demonstrates that head mixup

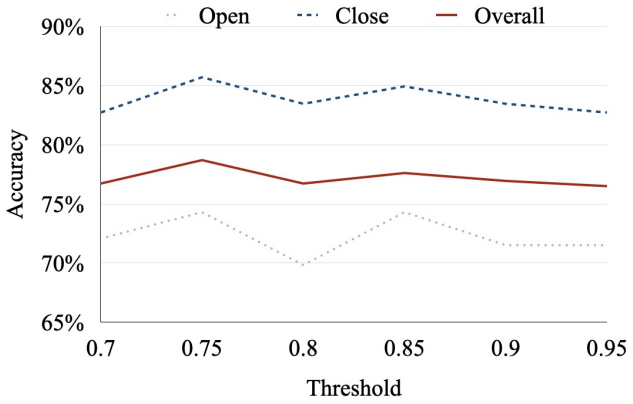


Figure 14: The Value of the Cosine Similarity Threshold in Difference Reconstruction.

provides data augmentation and decision boundary regularization through sample interpolation, effectively alleviating overfitting under the condition of limited medical data. Meanwhile, difference reconstruction encourages the model to reconstruct differences from similar images, highlighting subtle discriminative features. A balanced weighting of these two losses ensures that they complement each other and work synergistically to improve model performance.

To perform difference reconstruction, similar image pairs are identified by exploiting the pre-trained model's generalized ability to measure image similarity. This is achieved by extracting a global feature representation for each image and then computing the cosine similarity between image pairs. To identify the optimal similarity threshold, we conducted a thorough threshold sensitivity analysis, sweeping the threshold from 0.7 to 0.95 in increments of 0.05. As shown in Figure 14, the model achieves the best overall performance on downstream tasks when the cosine similarity threshold is set to 0.75.This outcome indicates that a threshold of 0.75 strikes an effective balance. It can filter out highly redundant samples with excessive visual similarity while preserving sufficient diversity. As a result, the difference reconstruction process can focus more precisely on discriminative visual differences, ultimately improving the model's generalization ability and downstream task performance.

In addition to using random sampling from a uniform distribution to determine the mixing coefficient λ in Equation (11), we also compared two alternative sampling strategies: Beta distribution sampling and fixed-value sampling, and evaluated their impact on the final model performance. Table 5 shows the results of Beta distribution sampling. This method tends to bias λ toward a specific interval depending on the chosen parameters, which limits the diversity of mixing ratios and makes it difficult for the model to cover a wide range of interpolated samples. Consequently,

Table 5: Comparison of Beta Distribution of the Mixing Coefficient λ in the Head Mixup.

Beta		**Open**	**Close**	**Overall**
α	β			
0.2	0.2	73.2%	82.4%	76.9%
0.4	0.4	73.7%	84.6%	77.6%
2.0	2.0	74.3%	83.5%	76.5%
0.2	1.0	76.5%	84.2%	78.1%
1.0	0.2	72.1%	83.8%	76.7%
Uniform (Ours)		74.3%	85.7%	**78.7%**

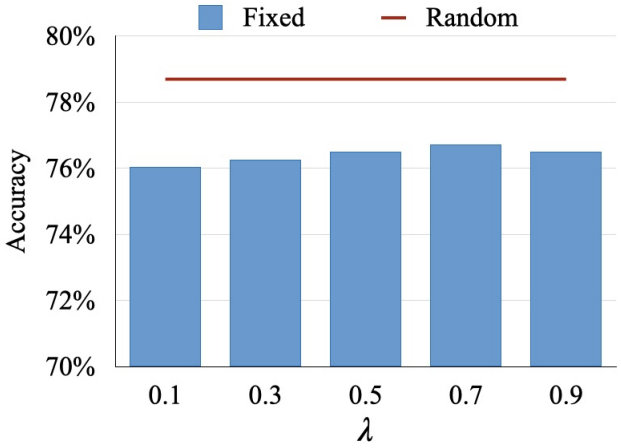


Figure 15: Comparison of Fixed-value Sampling of the Mixing Coefficient λ in the Head Mixup.

the regularization effect is weakened and the generalization ability is reduced.

The results of fixed-value sampling are shown in Figure 15. When λ is set as a constant, there is a lack of variability and randomness is completely eliminated, causing the model to overly rely on deterministic interpolation. This predictability restricts the effectiveness of data augmentation in combating overfitting under data-scarce scenarios and makes it difficult to simulate sufficient sample diversity. In contrast, random mixing with a uniform distribution introduces moderate randomness, which helps the model learn more robust multimodal fusion capabilities.

Notation Table

Table 6 offers a consolidated overview of the main abbreviations, symbols, and variables utilized in the paper, particularly those appearing in Section 3 and its related equations, to enhance clarity and readability. For comprehensive definitions of each symbol, please consult the relevant subsections within Section 3.

Table 6: Notation Table

Symbol / Abbreviation	Description	Reference
Key Abbreviations		
CMDPI	Cross-Modality Discriminative Pattern Identification Model	Abstract
VQA	Visual Question Answering	Abstract
LLM	Large Language Model	Section 1
MHA	Multi-Head Attention	Eq. 3
FFN	Feed-Forward Network	Eq.3
LN	Layer Normalization	Eq. 3
MSE	Mean Squared Error	Eq.5
GELU	Gaussian Error Linear Unit (activation in FFN)	Eq. 3
Feature Representations		
h_I	Global feature representation of the input image	Eqs.1 , 8
H_I	Local feature representations of the input image (patch sequence)	Eqs.1 , 8
h_T	Global feature representation of the input text	Eq.1
H_T	Local feature representations of the input text	Eqs. 1 ,8
h_F	Fused global representation	Eqs.1 , 2 , 8
$h'_{I/T}$	Masked global representations (image/text)	Eq. 2
$H'_{I/T}$	Masked local representations (image/text)	Eq.2
$\hat{h}_{I/T}$	Global representations from momentum encoders (contrastive learning)	Eq.6
$\hat{H}_I$	Local features of a similar image	Eq. 8
$\tilde{H}_I$	Deviated local image features (mix of original and similar)	Eq.8
$\tilde{h}$	Shared trainable vector for local reconstruction	Eq. 2
H_o^g	Hidden representations from global reconstruction	Eq. 2
H_o^l	Hidden representations from local reconstruction	Eq. 2
H_o^d	Hidden representations from difference reconstruction	Eq.8
h_D	Downstream head for answer prediction	Eq.11
h_{mix}	Mixed downstream head (for head mixup)	Eq.11
v_l	Intermediate representations in the reconstruction network	Eq. 3
$\hat{v}_l$	Normalized intermediate representations after MHA	Eq. 3
Loss Functions and Components		
L_{rec}	Overall reconstruction loss	Eq.5
L_{img}	Image reconstruction loss (MSE)	Eq. 5
L_{txt}	Text reconstruction loss (Cross-Entropy)	Eq.5
L_{ans}	Answer reconstruction loss (Cross-Entropy)	Eq.5
C_{inter}	Inter-modality contrastive loss	Eq.6
C_{intra}	Intra-modality contrastive loss	Eq. 6
C_{total}	Total contrastive loss	Eq. 6
L_{diff}^d	Difference reconstruction loss	Eq.10
L_{img}^d	Difference image reconstruction loss	Eq.10
L_{txt}^d	Difference text reconstruction loss	Eq.10
L_{mix}	Head mixup loss (Cross-Entropy)	Eq.12
L_{ft}	Overall fine-tuning loss	Eq. 13
Hyperparameters and Other Symbols		
ρ	Mask ratio for input masking	Eq.2
ϵ	Random factor for feature deviation (sampled from [0,1])	Eq.8
λ	Mixing coefficient for head mixup (sampled from [0,1])	Eq.11
α, α'	Scaling factors for inter- and intra-modality contrastive losses (0.5 and 1)	Eq.6
τ	Temperature parameter in contrastive loss	Eq.7
M	Number of text tokens	Eq. 4
$\text{sim}(a, b)$	Cosine similarity between vectors a and b	Eq.7
$\text{Contra}(a, b)$	Contrastive loss between representations a and b	Eq.7