Counterfactual Causal-Effect Intervention for Interpretable Medical Visual Question Answering

Linqin Cai, Member, IEEE, Haodu Fang, Nuoying Xu, and Bo Ren

***Abstract*—Medical Visual Question Answering (VQA-Med) is a challenging task that involves answering clinical questions related to medical images. However, most current VQA-Med methods ignore the causal correlation between specific lesion or abnormality features and answers, while also failing to provide accurate explanations for their decisions. Moreover, VQA-Med methods suffer from the common language bias problem in generic VQA. To explore the interpretability and language bias of VQA-Med, this paper proposes a novel CCIS-MVQA model for VQA-Med based on a counterfactual causal-effect intervention strategy. This model consists of the modified ResNet for image feature extraction, a GloVe decoder for question feature extraction, a bilinear attention network for vision and language feature fusion, and an interpretability generator for producing the interpretability and prediction results. The proposed CCIS-MVQA introduces a layer-wise relevance propagation method to automatically generate counterfactual samples for improving interpretability and alleviating language bias. Additionally, CCIS-MVQA applies counterfactual causal reasoning throughout the training phase to enhance interpretability and generalization. Extensive experiments on three benchmark datasets show that the proposed CCIS-MVQA model outperforms the state-of-the-art methods. Enough visualization results are produced to analyze the interpretability and debasing performance of CCIS-MVQA.**

***Index Terms*—Medical visual question answering, interpretability, counterfactual, causal-effect intervention.**

I. INTRODUCTION

ADVANCEMENTS in deep learning have successfully achieved state-of-the-art (SOTA) results in computer vision, natural language processing, and information retrieval. In the medical industry, deep learning technology has facilitated many significant applications. For example, several compelling studies have emerged in natural language processing using patient clinical records for predictive analysis [1], [2].

Visual Question Answering (VQA) [3] is a complex task in computer vision and natural language processing that aims to answer natural language questions relevant to given images. In the generic domain, deep learning has achieved great success with VQA. The migration of generic VQA to the medical field gives rise to a new downstream task: Medical Visual Question Answering (VQA-Med). In the VQA-Med task, radiological scans of the patients (X-ray, Magnetic Resonance Imaging (MRI), and CT) are used instead of standard images in the generic domain, accompanied by clinically relevant question-answer (QA) pairs. VQA-Med technology can assist doctors in improving the diagnosis efficiency and help patients understand their conditions. However, VQA-Med is challenging as it demands an in-depth understanding and high-level interactions with professional medical images and textual QA pairs to generate reasonable and credible answers.

Early methods in the field of VQA-Med have attempted to fine-tune the deep networks by using existing VQA models on generic VQA datasets. For example, Bansal et al. [4] used a ResNet image embedding and word embedding from a pre-trained Word2Vec model on the PubMed dataset to generate descriptions of abnormalities in the images. Vu et al. [5] used a pre-trained ResNet-152 model on the ImageNet dataset to extract image features and employed the pre-trained bidirectional encoder representations from transformers (BERT) to extract question features. They also proposed a VQA approach that leveraged a bilinear model to aggregate and synthesize the extracted image and question features. However, these methods only used low-level features of both structural questions and non-structural images, ignoring the causal correlation between specific lesion or abnormality features and answers in medical images. This approach also fails to provide appropriate explanations for predictions that are understandable to humans. Interpretability is critical to producing convincing answers for the reliability and trustworthiness of VQA-Med to help doctors understand patients comprehensively and make correct and appropriate clinical decisions. Furthermore, most VQA-Med methods suffer from the common language bias problem found in generic VQA, as they often rely on spurious textual cues to make decisions rather than engage in multi-modal reasoning. The linguistic prior process of training QA pairs in unbalanced training datasets also seriously affects these models.

Capturing causal correlation between image and QA pairs is crucial to enhance features from image models and achieve

This work was supported by the National Natural Science Foundation of China under Grant 62277008.

Linqin Cai is with the Research Center for Artificial Intelligence and Smart Education, Chongqing University of Posts and Telecommunications, Chongqing, 40065, China (e-mail: iamlqcai@163.com).

Haodu Fang is with the Research Center for Artificial Intelligence and Smart Education, Chongqing University of Posts and Telecommunications, Chongqing, 40065, China (e-mail: fhdu1219@163.com).

Nuoying Xu is with the Research Center for Artificial Intelligence and Smart Education, Chongqing University of Posts and Telecommunications, Chongqing, 40065, China (e-mail: xunuoying2021@163.com).

Bo Ren is with the Research Center for Artificial Intelligence and Smart Education, Chongqing University of Posts and Telecommunications, Chongqing, 40065, China (e-mail: 3244498655@qq.com).

interpretability and good debiasing performance for VQA. In recent years, counterfactual reasoning has become one of the promising pipelines for interpretability in explainable artificial intelligence [6]. Counterfactual reasoning provides an interpretation at the level of human knowledge by answering the question of "*What does X have to change to alter the prediction from Y to Y'?*" and explaining a model's decision in hypothetical scenarios. From a perspective of causality [7], humans learn by actively interacting with the environment and infer causal dependencies between events by intervening and observing changes in the outcomes. Therefore, causal relationships can affect the final target results through human intervention [8]. According to this premise, Wang et al. [9] proposed a causal-effect intervention strategy (CIS) based on interpretable vision models to proactively guard against image features with no causal relevance for improving image classification performance and facilitating model interpretability. Niu et al. [10] proposed a counterfactual inference framework (CF-VQA) to capture the language bias as a direct causal effect of questions on answers and to reduce the language bias in VQA. Chen et al. [11] presented a more sophisticated counterfactual samples synthesizing training scheme to enhance the visual interpretability and question sensitivity of VQA. Inspired by this philosophical concept of counterfactual reasoning [8], image classification [9], and the generic VQA [10], [11], this study aims at constructing a novel counterfactual causal-effect intervention strategy (CCIS) framework for VQA-Med to explore the problems of interpretability and language bias of VQA-Med. Through contrastive language-image pre-training (CLIP), we apply the modified ResNet and transformer decoder to capture low-level image features and question features, respectively. Then, we fuse the vision and language features using a bilinear attention network (BAN) to obtain high-level global features. Building upon this, we introduce a layer-wise relevance propagation (LRP) method to automatically generate counterfactual image training samples that enhance the model's interpretability and alleviate the language bias from the VQA-Med model and datasets. Unlike other methods of generating counterfactual samples with artificially specified rules, our proposed causal intervention strategy can simultaneously produce interpretability and prediction results. Throughout the training phase, we apply the structural causal model (SCM) [8] based on counterfactual causal reasoning to further enhance the interpretability and prediction performance of the VQA-Med model. This study provides three-fold contributions.

First, we propose a CCIS framework for VQA-Med (CCIS-MVQA) to explore the interpretability and language bias of VQA-Med, which integrates an interpretability generator into its architecture to provide interpretations and explanations for its predictions.

Second, we develop layer-wise relevance propagation to automatically generate counterfactual image samples and construct a CCIS strategy, which can simultaneously produce interpretability and prediction results.

Third, we perform extensive experiments on benchmark datasets. Our proposed CCIS-MVQA achieves new SOTA results compared with the existing methods in VQA-Med fields. Additionally, we provide sufficient visualization results to analyze the interpretability and debasing performance.

II. Related Works

This paper explores the problems of interpretability and language bias in VQA-Med. Here we begin by discussing language bias and reviewing related to the VQA-Med interpretability.

In the VQA model, language bias refers to giving the predicted answer based on spurious linguistic correlations between questions and training data without comprehensively reasoning multi-modal information from the images and texts because of an unbalanced training dataset [10], [12]. Most current solutions to reduce language bias in VQA fall into three categories: enhanced visual grounding [13], [14], weakened linguistic priors [10], [12], and explicit/implicit data balancing [11], [15]. CF-VQA [10] and RUBi [12] are two well-known VQA debias methods in generic VQA. CF-VQA [10] is a counterfactual inference framework for capturing language bias. However, the counterfactual samples were synthesized according to artificially specified rules. Cadene et al. [12] proposed reducing unimodal biases for VQA (RUBi). It used additional QA models to capture language bias, whereas the QA models were not used in the test stage.

Saliency mapping is a popular method for interpreting deep learning models for the interpretability of generic VQA. Wang et al. [9] have presented a proactive pseudo-intervention strategy that proactively guards against image features with no causal relevance. Zhang et al. [16] have proposed generating a significant heat map to display the image regions related to the answers. However, Fernandez et al. [17] pointed out that the above methods did not adequately explain VQA decision-making. Teney et al. [18] have introduced masking the overlap between the boundary box and the artificially annotated attention map to enhance visual interpretability. Pan et al. [19] have also proposed generating counterfactual images by editing the original images. However, due to the complexity of the questions in VQA-Med, this framework can only be used for color-related questions. Most existing research uses feature-based post-hoc interpretation to describe the decision-making process they want to explain. Existing VQA-Med models tend to employ some advanced methods used in generic VQA, such as multi-modal compact bilinear [20], stacked attention networks [21], bilinear attention networks [22], multi-modal factorized bilinear [23], and multi-modal factorized high-order [24].

Various approaches have been developed to address the challenge of limited labeled data in the VQA-Med task. Nguyen et al. [25] proposed a mixture of enhanced visual features (MEVF) model that uses meta-learning to adapt to VQA-Med tasks with limited labeled data, making the model effectively learn meta-annotations [26]. Liu et al. [27] introduced a contrastive pre-training and representation distillation (CPRD) to train teacher models using many unlabeled radiological images. Then, these models were transferred into a lightweight student model for fine-tuning radiological images of VQA-Med

datasets. Another study [28] proposed a bi-branched model based on parallel networks and image retrieval for VQA-Med (BPI-MVQA) to realize complementary advantages in image sequence feature extraction, spatial feature extraction, and multi-modal fusion, forcing the VQA-Med to consider the feature applicability to specific image-understanding tasks [29].

Various proposals have been used to simplify the complex task of medical VQA. Ren et al. [30] introduced a classification and generative model for VQA-Med (CGMVQA), which uses a multi-head self-attention mechanism to break the complex medical VQA task into multiple simple tasks. Zhan et al. [31] proposed a conditional reasoning framework for various VQA-Med tasks to automatically learn practical reasoning skills for various VQA-Med tasks. To emphasize question features, Vu et al. [32] developed a question-centric multi-modal low-rank bilinear (QC-MLB) model for VQA-Med to fuse image and question features by enforcing high adherence to the query sentence. Zhang et al. [33] and Pan et al. [34] improved the reasoning ability of different models using attention mechanisms to realize the feature alignment based on text and image and enhance the semantic alignment ability of cross-modal features. Yu et al. [35] proposed a question-guided feature pyramid network (QFPN) for VQA-Med, using high resolution of low-level features and rich semantic information of high-level features to capture multi-scale information of medical images.

Some studies have explored question features in the VQA-Med method. Cong et al. [36] proposed a technique called Caption-Aware, which used caption-sensing to understand the abstract information of image content and clinical diagnosis from many medical images. Li et al. [37] introduced a bi-level representation learning model for VQA-Med using sentence- and word-level reasoning. Huang et al. [38] proposed a VQA-Med network based on medical knowledge to learn the disease-related and relation-related embedding according to the structural features of a medical knowledge graph. Cong et al. [39] developed an anomaly-oriented model (AOM) based on weakly supervised anomaly localization information using generative adversarial networks to generate healthy images and anomaly localization result maps. This paper also uses abnormal location information and irrelevant information in input images. However, we still use causal correlation to reason QA pairs effectively. Additionally, we quantify the model's debias ability.

Existing works [31] in VQA-Med primarily apply advanced methods in generic VQA. However, medical data image pattern and linguistic style differ significantly from the generic domain. Compared to general VQA, VQA-Med task requires higher-level reasoning skills such as locating specific lesions or evaluating if the size of an organ is expected relative to prior knowledge. Additionally, questions in VQA-Med must be more realistic and specific, making it harder to collect or generate these questions [39]. As a result, it is complicated to obtain well-annotated datasets for training VQA-Med systems. Due to these differences between generic VQA and VQA-Med, simply generic VQA methods and fine-tuning limited medical data provide little benefit [31].

III. Method

The VQA-Med task is a multi-class classification problem, and we consider an image-question pair (V,Q), where Q represents a medical-related question, and V is a medical image. A dataset $D=\{v_i,q_i,a_i\}^{N_i}$ consists of triplets of images $v_i \in V$, questions $q_i \in Q$, and candidate answers $a_i \in A$. VQA-Med can be expressed as a question-and-answer model to find the answer with the highest probability from candidate answers as:

$$\hat{a}=\arg\max_{a_i \in A} P_\theta\left(a_i \mid v_i, q_i\right), \tag{1}$$

where θ is the parameter in question answering model.

As shown in Fig. 1, the proposed CCIS-MVQA consists of four main components.

1) Image feature extraction uses modified ResNet with convolutional neural network to capture low-level visual features.

2) Question feature extraction uses GloVe and transformer decoder to capture contextual information.

3) Bilinear attention network fuses the vision and language features for high-level global features.

4) Interpretability generator applies CCIS and causal reasoning to enhance the model's interpretability and generalization. The image encoder uses modified ResNet50 for the backbone of the CCIS-MVQA model. However, the text encoder uses a transformer and GloVe for the same model, where the transformer pre-trained the network using contrastive language-image pre-training CLIP [40], [41], and GloVe is for fine-tuning.

A. Image Feature Extraction

This paper uses modified ResNet50 to extract and enhance image features, especially fine-grained information. We fine-tune the modified ResNet50 [42] based on the original CLIP model [40], [41] pre-trained on a public domain image-text dataset. We have also conducted this fine-tuned process using medical image-text pairs on the ROCO dataset [43], which contains various image forms (X-rays, ultrasounds, and MRIs) for almost all body parts. Moreover, each image has corresponding descriptive text to provide explanatory information.

Given N image-text pairs, we train CLIP to classify $N \times N$ image-text combinations for CLIP fine-tuning task and VQA-Med. In this process, CLIP learns image and text embedding features. It optimizes the model parameters by maximizing the cosine similarity of the image and text embedding during the gradient calculation phase. For each batch, the expected prediction is in a two-dimensional matrix, and the number of false image-text pairs is N^2-N. The similarity score is optimized using a symmetric cross-entropy loss function.

The modified ResNet50 differs from the original ResNet50. The input stem of the modified ResNet50 replaces one convolutional layer of ResNet50 with three convolutional layers and replaces the maxpool layer with the global average pool layer. Additionally, the downsampling block in the modified ResNet50 replaces the first convolutions layer (1 × 1,

s = 2) with (1 × 1, s = 1). Finally, the modified ResNet50 replaces the last linear classification header with a self-attention layer and calculates the vectors through a linear mapping matrix. Based on the model, we reshape an image I to match ResNet50 (224, 224, 3), replace the global average pooling layer with an attention pooling mechanism, and retain only the output of the last average pooling layer as image features X_I as:

$$X_I = \text{ResNet50}(I) . \tag{2}$$

Attention pooling is implemented as a single layer of a "Transformer-style" multi-head attention mechanism, where the query is conditioned on the global average pooling representation as:

$$Q = W_Q X_I, \quad K = W_K X_I, \quad V = W_V X_I$$
$$X = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V , \tag{3}$$

where W_Q, W_K, and W_V are attention mapping weights of X_I.

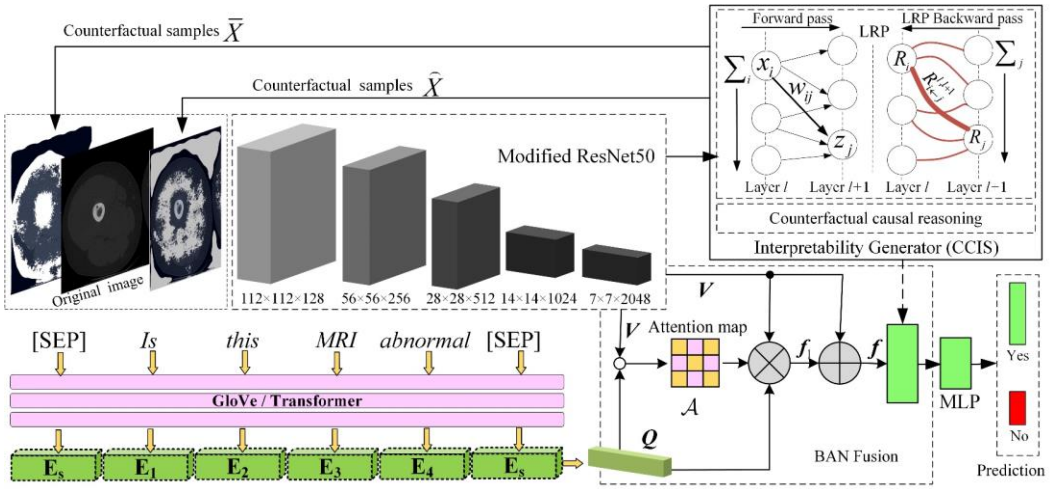


Fig. 1. The proposed CCIS-MVQA framework based on counterfactual causal-effect intervention strategy

B. Question Feature Extraction

The process of feature extraction for questions involves two stages: the pre-training stage and the fine-tuning stage. We use a transformer decoder to extract question features during the pre-trained stage. As a base size, the question feature network is a 12-layer attention network of 512 widths and 8 heads. The text sequence is marked with [SOS] and [EOS] and activated at the highest level of the transformer as a feature representation of questions. Subsequently, text features are layered, normalized, and linearly projected into embedding space. In the fine-tuning stage, questions are encoded by GloVe to obtain text vectors $\boldsymbol{Q} \in \mathbb{R}^{M \times d_q}$, where M is the length of the text sequence, and d_q is the dimension of $\boldsymbol{Q}$. However, the transformer in CLIP trained with 400M image-text pairs should be better than Glove. This paper used the CLIP to enhance the image modality. Therefore, the Glove is deliberately used to save computing time and indirectly strengthen the image modality to weaken the impact of language modality.

C. Multimodal Fusion

In the feature fusion stage, the modified ResNet50 pre-trained by CLIP is used as an image feature extractor to obtain fine-grained image features $\boldsymbol{V} \in \mathbb{R}^{N \times d_v}$, where N is the image sequence's length and $\boldsymbol{d}_v$ is the dimension of V. GloVe encodes text features to obtain text vectors $\boldsymbol{Q} \in \mathbb{R}^{M \times d_q}$, where M is the length of the text sequence, and d_q is the dimension of $\boldsymbol{Q}$. The obtained image features and text features are fused through bilinear attention network as:

$$\boldsymbol{f} = \sum_i^K \left(\boldsymbol{V}^T \boldsymbol{W}_v\right)_i^T \mathcal{A} \left(\boldsymbol{Q}^T \boldsymbol{W}_q\right)_i^T + \left(\mathbf{I}^T \bullet \boldsymbol{V}^T\right) \boldsymbol{W}_V , \tag{4}$$

where $f \in \mathbb{R}^K$ is the cross-modal vector; K is the dimension of f, $\boldsymbol{W}_v \in \mathbb{R}^N$, $\boldsymbol{W}_q \in \mathbb{R}^M$ is a learnable weight parameter matrix, N and M are the dimensions of the matrix, and $\mathcal{A} \in \mathbb{R}^{d_v \times d_q}$ is a bilinear attention map used to fuse the image and text features as:

$$\mathcal{A} = \text{softmax}\left(\left(\left(\mathbf{I} \cdot \boldsymbol{p}^T\right) \circ \boldsymbol{V}^T \boldsymbol{W}_V\right) \boldsymbol{W}_Q^T \boldsymbol{Q}\right), \tag{5}$$

where $\boldsymbol{I} \in \mathbb{R}^{d_v}$ is used to change the shape of bilinear attention maps; $\boldsymbol{p} \in \mathbb{R}^{1 \times K}$, $\boldsymbol{W}_V \in \mathbb{R}^{N \times K}$, and $\boldsymbol{W}_Q \in \mathbb{R}^{M \times K}$ are weight parameter matrixes of bilinear attention network, respectively. Afterward, a multilayer perceptron classifier (MLP) completes answer predictions.

D. Counterfactual Causal-effect Intervention Strategy

When people make causal reasoning on the surrounding events, they often have the retrospective counterfactual thinking that "*What would have happened had we acted differently*?" [7], [8]. Answering a counterfactual question allows us to learn from history and the experience of others. As a unique expression to causal reasoning [8], counterfactual explanations can provide a more sophisticated way to increase the interpretability of VQA model [11], [17].

In one study [8], Pearl divided causality into three levels: *Association*, *Intervention*, and *Counterfactuals* from bottom to top. At the bottom level, the *Association* is to find correlations

between variables from the observed data. The second level *Intervention* shows that when changing X, whether Y will change with X. At the top level, *Counterfactuals* indicate that if we want Y to undergo a specific change, we can achieve it by changing X. As a result, we aim to study the interpretability of the VQA-Med model at the *Intervention* and *Counterfactuals* rungs and propose a new CCIS.

(1) Interventions

A true causality can be intervened to influence the outcomes. Formally, intervention can be explained with *do*-notation [7] as:

$$P(Y \mid do(x)) = P(Y \mid X = x), \quad (6)$$

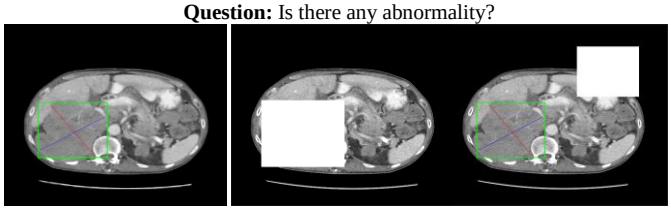
where we identify Y as the answer and X as the image. In (6), variable X is artificially forced to take a value x, but otherwise, the remaining variables are stimulated according to the original process of generating data to study the changes in the distribution of Y.

However, performing an interventional study in reality, i.e., randomized control trials, intentionally blocking non-causal associations is often not feasible due to cost, time, and ethics [6], [9]. This study applies synthetic interventions to uncover the underlying causal features from observational data by automatically editing image X and its corresponding label Y to encourage the VQA-Med model and explore potential causal interpretability.

(2) Counterfactuals

A counterfactual statement might be interpreted as conveying a set of predictions under a well-defined set of conditions that prevail in the factual part of the statement [7]. More precisely, counterfactuals have expressions of the type $P(Y_x \mid \bar{X}, \bar{Y})$, which stand for "the probability that event $\gamma = Y$ would be observed had χ been X, given that we actually observed X to be $\bar{X}$ and Y to be $\bar{Y}$ " [44]. Fig. 2 shows a counterfactual interpretation for VQA-Med by intervention to mask a critical region of the input image and thus change the answer distribution.

Question: Is there any abnormality?



Prediction: Yes **Prediction:** No **Prediction:** Yes

(a) Before intervention (b) After intervention

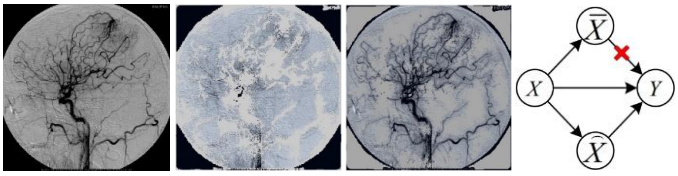
Fig. 2. Factual and counterfactual interpretations on VQA-Med dataset.

Before intervention in Fig. 2(a), the VQA-Med model receives a complete anomaly image and predicts the correct answer according to the abnormal location information, indicating an abnormal lesion in the image. The answer to "*Is there any abnormality?*" should be "Yes". After intervention in Fig. 2(b), if the abnormal lesion parts in the image are masked, VQA-Med model will not predict the correct answer, i.e., the answer should change from "Yes" to "No" (left); even if other parts of the image other than the abnormal lesion parts are masked, the answer should still be "Yes" (right).

(3) Structural Causal Model

The structural Causal Model (SCM) [8] defines the data-generating process and the distribution of the observations. For causal-effect reasoning, we apply the SCM model [8] to elaborate causal relationships between the original sample X, counterfactual samples $\bar{X}$ and $\hat{X}$, and answer labels Y, respectively. Fig. 3 clearly shows how the original sample X and counterfactual samples $\bar{X}$ and $\hat{X}$ use a reasoning path to reason about causal relationship with the answer label Y.

Question: What's the primary abnormality?
Answer: Vascular malformation



(a) X (b) $\bar{X}$ (c) $\hat{X}$ (d) SCM model

Fig. 3. Causal graph model based on counterfactual samples. (a) Original image X; Answer Y: Vascular malformation. (b) Counterfactual image $\bar{X}$; Answer Y: **NOT** Vascular malformation. (c) Counterfactual image $\hat{X}$; Answer Y: Vascular malformation. (d) SCM model.

Based on the original sample X, we perform a pixel-wise mask on X to generate counterfactual images $\bar{X}$ and $\hat{X}$. Counterfactual samples $\bar{X}$ are generated by masking the critical parts of the original image, in which these covered objects are essential and casually correlated to correctly answering a question, e.g., the gray pixel masks on the blood vessels region in Fig. 3(b). Counterfactual samples $\hat{X}$ are introduced as the adversarial control group and may not adversely affect the correctly answering the question. $\hat{X}$ is generated by masking the non-critical parts of the original image, in which these covered objects are not casually correlated to correctly answering a question, e.g., the gray pixel masks at the edges of the image other than the blood vessels in Fig.3(c). Y is the ground-truth answer. Then, the counterfactual images and original questions compose a new visual image-question (VQ) pair.

Given a VQ pair containing counterfactual image samples, a standard VQA training sample triplet still needs the corresponding ground-truth answers. For X and $\hat{X}$, the ground-truth answer already exists in the dataset. The counterfactual image $\bar{X}$, which masks the critical parts of the original image, will affect the prediction answer, so the ground-truth answer must be different from the original one. We introduce a layer-wise relevance propagation technique to automatically generate counterfactual samples, aiming to avoid expensive manual annotations.

As shown in Fig. 3(d), a connection represents the causal relationship between two nodes: cause $\rightarrow$ effect. There are three causal-effect reasoning paths in Fig. 3(d). (i) $X \rightarrow Y$, given a VQ pair that contains the original image, the VQA-Med model can predict a correct answer based on the critical information in the image if the model is accurate enough. In this case, the VQA-Med model can capture the direct causal effect of the image. (ii) $(X \rightarrow \hat{X} \rightarrow Y)$, given a VQ pair that contains a counterfactual image with non-critical parts masked, the VQA-

Med model can still predict a correct answer based on the critical parts of the image if the model is accurate enough. (iii) $\left(X \to \bar{X} \xrightarrow{\text{NOT}} Y\right)$, given a VQ pair that contains a counterfactual image with critical parts masked, VQA-Med should not predict the correct answer even if it is accurate enough.

Fig. 3 clearly shows that regardless of whether the image pixel is related to the prediction, using counterfactual samples, the causal path in the SCM model can help the VQA-Med model correlate original samples X with answer labels Y. According to the causal-effect reasoning paths of SCM models in Fig. 3, this paper introduces a CCIS to eliminate irrelevant causal relationships so that the VQA-Med model can focus on the critical objects or regions with causal correlations in the input image. Adding an intervention strategy for the image modal into the CCIS-MVQA model can improve interpretability and alleviate the deviation brought by the language modal. On this basis, CCIS-MVQA constructs a loss function to train and optimize model parameters, which will be discussed in detail later.

(4) Counterfactual sample generation

We introduce a layer-wise relevance propagation (LRP) method [45], [46] for generating counterfactual samples. The contribution of each pixel in the image to model predictions is calculated through the layer-wise relevance backpropagation in the image feature extraction network. According to the causal relationship, the region with a larger contribution value can better determine the prediction answer, called the causal saliency map [9]. Therefore, counterfactual samples $\bar{X}$ and $\hat{X}$ can be obtained by masking the pixels with largest or smallest contribution values.

The core idea of LRP is to decompose the target function into a set of correlation scores and then redistribute them to the neurons in the previous layer, as shown in the interpretability generator module on the upper right of Fig. 1. The rule of correlation scores R_j propagating from the (l+1)-th layer to the previous l-th layer is in (7).

$$R_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j^{(l+1)}, \tag{7}$$

where i and j are the neuron nodes of the adjacent layer, respectively; $R_i^{(l)}$ is the correlation score of the i-th neuron in the l-th layer, and $R_j^{(l+1)}$ is the correlation score of the j-th neuron in the (l+1)-th layer, respectively; Σ is the sum of the correlation scores of all nodes; a_i is the attribute of the *i-th* node; w_{ij} is the weight from the i-th neuron to the j-th neuron. $a_i w_{ij}$ quantifies the contribution of node i to node j in the forward propagation process. The denominator in (7) ensures the conservation of propagation, indicating that the information received by the neurons must be reallocated equally to the next layer.

As shown in Fig. 1, the interpretability generator obtains the causal saliency map $s(x_i)$ of the prediction answer using LRP, giving an input image X and obtaining the image feature x_i by image feature extraction. We mask the critical information (image area with largest contribution value) for counterfactual images as:

$$\bar{x}_i = x_i - F\left(s\left(x_i\right)\right) \odot x_i, \tag{8}$$

where $F(\cdot)$ is a mask function as:

$$F\left(s\left(x_i\right)\right) = \frac{1}{1+\exp\left(-k\left(s\left(x_i\right)-\sigma\right)\right)}, \tag{9}$$

where k is a threshold parameter for controlling the mask range; σ is a scaling parameter for specifying the mask color; $\sigma > \boldsymbol{0}$. By selecting salient pixels, parameter k is designed to encourage the causal parts of each image to be as small as possible to avoid possible degeneration solutions of the objective function.

In addition, we should avoid interference resulting from the intervention strategy because the VQA-Med model does not learn to capture causal correlation but learns the intervention operation (masking image). For example, a VQA-Med model can learn to change its predictions when it detects that an input image is masked regardless of whether the image lacks causal features, which may harm the predicted outcomes. Therefore, an adversarial control group, i.e., a counterfactual sample $\hat{x}$, is introduced to mask the non-critical parts of the original images. The counterfactual sample $\hat{x}$ is generated as:

$$\hat{x}_i = x_i - F\left(x_i - s\left(x_i\right)\right) \odot x_i. \tag{10}$$

Notably, (10) may lead to degenerate solutions; that is, any counterfactual sample generated by a causal salient map satisfying causal-effect correlation is an effective mask regardless of the covered area size. For example, covering the whole image or only covering a lesion removes the causal-effect correlation between the critical region of the image and the answer. This is detrimental to alleviating model language bias and improving model interpretability. The k in (9) is used to control mask range to avoid this issue, which encourages the masked parts to be only a small part of the image.

(5) Model training and optimization

Counterfactual samples are integrated into the training process of the CCIS-MVQA model, as shown in Fig. 4. According to the causal-effect reasoning paths described in SCM models in Fig. 3, CCIS-MVQA constructs a loss function to participate in training and optimizing model parameters at each training epoch. Our loss functions are as follows:

$$L_i^1(\theta) = \ell\left(\bar{x}_i, q_i, \neg y; f_\theta\right), \tag{11}$$

$$L_i^2(\theta) = \ell\left(\bar{x}_i, q_i, \neg y; f_\theta\right), \tag{12}$$

$$L_i^3(\theta) = \ell\left(x_i, q_i, y; f_\theta\right), \tag{13}$$

$$Loos_i = L_i^1 + L_i^2 + L_i^3, \tag{14}$$

where f_θ is the prediction function of the VQA-Med model; θ is the model parameter; y is the ground-truth answer, $\neg y$ denotes a flip of ground-truth answer, i.e., $\ell\left(\bar{x}_i, \neg y; f_\theta\right) = -\ell\left(\bar{x}_i, y; f_\theta\right)$. In this way, we do not need to add additional ground-truth answers.

As shown in Fig. 4, CCIS-MVQA applies CCIS to integrate counterfactual samples into the model training process and constructs loss functions to participate in the model parameters optimization. According to the CCIS causality, CCIS-MVQA

will predict a wrong answer when using counterfactual samples $\bar{x}$ and predict a right answer when using counterfactual samples $\tilde{x}$. As shown in Fig. 4, at each training epoch *i*, the interpretability generator in CCIS-MVQA can update the model parameters and network weights according to the loss function and automatically generate counterfactual samples more causally correlated with the prediction label *Y*. The updated parameters will be further optimized in the next training epoch *i*+1.

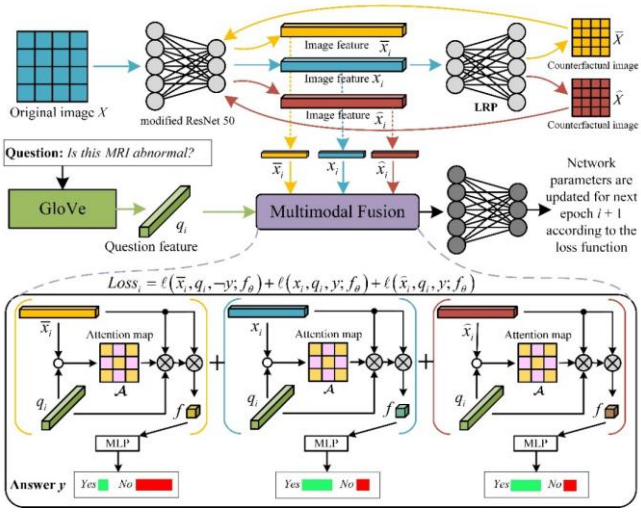


Fig. 4 Training and optimization process of CCIS-MVQA at epoch *i*.

After updating weight parameters at each training epoch, CCIS-MVQA can further generate more accurate counterfactual samples to improve causal reasoning and prediction performance continuously. This approach achieves a benign causal cycle, which captures the causal-effect interpretability by inferring the critical objects or regions in the image closely causally related to the prediction outcomes and enhances the explainability of the CCIS-MVQA model. This is also a self-explanatory approach in which CCIS-MVQA integrates the interpretability generator module into its architecture to explain its predictions.

IV. Experiments And Results Analysis

A. Datasets

This paper evaluates the proposed CCIS-MVQA model using three publicly available VQA-Med datasets: VQA-Med-2019 [47], VQA-RAD [48], and SLAKE [49]. In general, most images in the datasets match multiple pairs of questions and answers, which are divided into two types: closed-ended and open-ended. Closed-ended questions only answer "Yes" or "No" and open-ended questions are answered in free-form text

B. Implementation details

The proposed CCIS-MVQA model is implemented with the PyTorch library. The medical image model is trained from random initialization with the Adam optimizer. The initial learning rate is 0.001, the momentum is 0.05, and the batch size is 16. We create 300-D GloVe embeddings for question embedding. For the transformer encoder, the hidden size is $16\times16\times3$; the number of heads is 8, the batch size is 16, the number of epochs is 200, and the dropout rate is 0.5 for all layers.

C. Evaluation of the overall performance

We compare our CCIS-MVQA model with some SOTA methods: MEVF [25], CPRD [27], BPI-MVQA [28], CGMVQA [30], QC-MLB [32], QFPN [35], Caption-Aware [36], AOM [39], and Optimal Model [50]. We have reviewed these models in Section II.

This paper selects accuracy (i.e., *Acc.*) as the evaluation metric, the percentage of correctly predicted instances to the total predicted cases. This evaluation standard is the simplest and strictest, a new metric introduced in the ImageCLEF 2019 VQA-Med dataset [47], and strictly considers the exact matching of the predicted answer and the ground truth answer. Table I shows the evaluation results. The best results are highlighted with bold values in each column. The superscript values of our CCIS-MVQA are the Mean Square Error (MSE) variance estimates for the accuracies.

TABLE I

Quantitative Verification Results of Model Overall Performance (*Acc.* %)

Model	VQA-RAD			SLAKE		
	Opened	Closed	Overall	Opened	Closed	Overall
MEVF + BAN [25] (2019)	43.90	75.10	62.60	77.80	79.80	78.60
MEVF + SAN [25] (2019)	40.70	74.10	60.80	75.30	78.40	76.50
CPRD + BAN [27] (2021)	52.50	77.90	67.80	79.50	83.40	81.10
Caption-Aware [36](2022)	65.40	77.90	72.00	79.60	86.10	82.20
CCIS-MVQA (Ours)	**68.78**$^{\pm0.23}$	**79.24**$^{\pm0.16}$	**75.06**	**80.12**$^{\pm0.11}$	**86.72**$^{\pm0.07}$	**84.08**
Model	VQA-Med-2019					
	Modality	Plane	Organ	Abnormality		All
CGMVQA [30] (2020)	81.92	86.47	78.47	4.47		62.83
QC-MLB [32] (2020)	82.45	73.17	70.94	4.85		57.85
Optimal Model [50] (2022)	52.02	62.15	48.06	6.08		42.08
BPI-MVQA [28](2022)	84.83	84.80	72.81	**19.20**		65.41
AOM [39](2022)	55.15	86.75	68.55	-		-
QFPN [35](2023)	-	-	-	-		63.80
CCIS-MVQA (Ours)	**88.78**$^{\pm0.38}$	**88.16**$^{\pm0.21}$	**84.18**$^{\pm0.14}$	12.35$^{\pm0.40}$		**68.37**

In Table I, our CCIS-MVQA achieves 75.06% and 84.08% overall accuracy on the VQA-RAD and SLAKE, respectively. Compared with the best baseline Caption-Aware [36], CCIS-MVQA improves the overall accuracy by 3.06% and 2.88%, respectively. For the open and closed questions, CCIS-MVQA obtains the best results.

In the VQA-Med-2019 database, our CCIS-MVQA has achieved more competitive improvement and obtained SOTA results with a 68.37% average accuracy. Compared with the second-best baseline BPI-MVQA [28], our CCIS-MVQA increases average accuracy by 2.96%. Our CCIS-MVQA also obtains impressive results in the specific question categories except in the "Abnormality" category. Our CCIS-MVQA surpasses all models except the best BPI-MVQA in the abnormality category [28]. The "Abnormality" question type in VQA-Med-2019 presents a significant barrier to reasoning the answer because it contains more than 1000 categories of abnormality answers. BPI-MVQA [28] designed an individual branch for image retrieval to predict the irregular, open-ended 'Abnormality' type questions.

D. Interpretability evaluation:

This paper adopts the proposed evaluation scheme [51] to quantitatively evaluate the causal saliency map's causal correlation. As in [51], we compute an estimate e of feature importance for every input pixel in the dataset using the LRP method, and then rank each e into an ordered set $\{e_i^o\}_{i=1}^N$. For the top k fraction of the ordered set, we replace the corresponding pixels in the raw image with the grayscale value 255 to generate new datasets at different degradation levels k = [0, 10, 20, 30, and 40], where k is a mask range of the image in (9). Then, the CCIS-MVQA model is evaluated on the new test data.

Table II shows the decline in prediction accuracy after removing critical pixels. As the k increases, the prediction accuracy of the model decreases gradually, indicating that the proposed CCIS-MVQA captures the critical part of the image that has a causal correlation with the predicted answer, further reflecting that the proposed CCIS-MVQA model minimizes the use of irrelevant background information in decision-making.

TABLE II

EFFECT OF DIFFERENT K VALUE ON CCIS-MVQA PERFORMANCE (*Acc.* %)

Dataset \ k	10%	20%	30%	40%
VQA-RAD	-3.2	-12.6	-19.2	-20.2
SLAKE	-3.1	-11.4	-17.4	-18.9
VQA-Med-2019	-3.5	-9.2	-18.6	-19.4

Fig. 5 shows more intuitive examples of counterfactual samples with different k values, where the image's gray pixel masks are casual or non-casual correlated. In each two-line legend, the casual correlated counterfactual sample is at the top, and the non-casual correlated counterfactual sample is at the bottom. Each group's casually correlated counterfactual samples accurately placed the masks in the correct region, which casually correlated with correctly predicting an abnormality. Furthermore, the non-casual correlated counterfactual samples of each group place the masks in the irrelevant background position. With the increase in parameter k, the mask range gradually expands, which proves the effectiveness of the proposed causal intervention strategy.

As shown in the first row in Fig. 5(a), the gray pixel masks cover critical parts of the image that are important for correctly answering a question related to pneumocystis carinii pneumonia, the abnormal lung region. In contrast, in the second row, the gray pixel masks cover the background regions (e.g., the pixels at the edge of the image) that are unimportant for correctly answering a question related to pneumocystis carinii pneumonia, indicating that influential objects are more related to the QA pair.

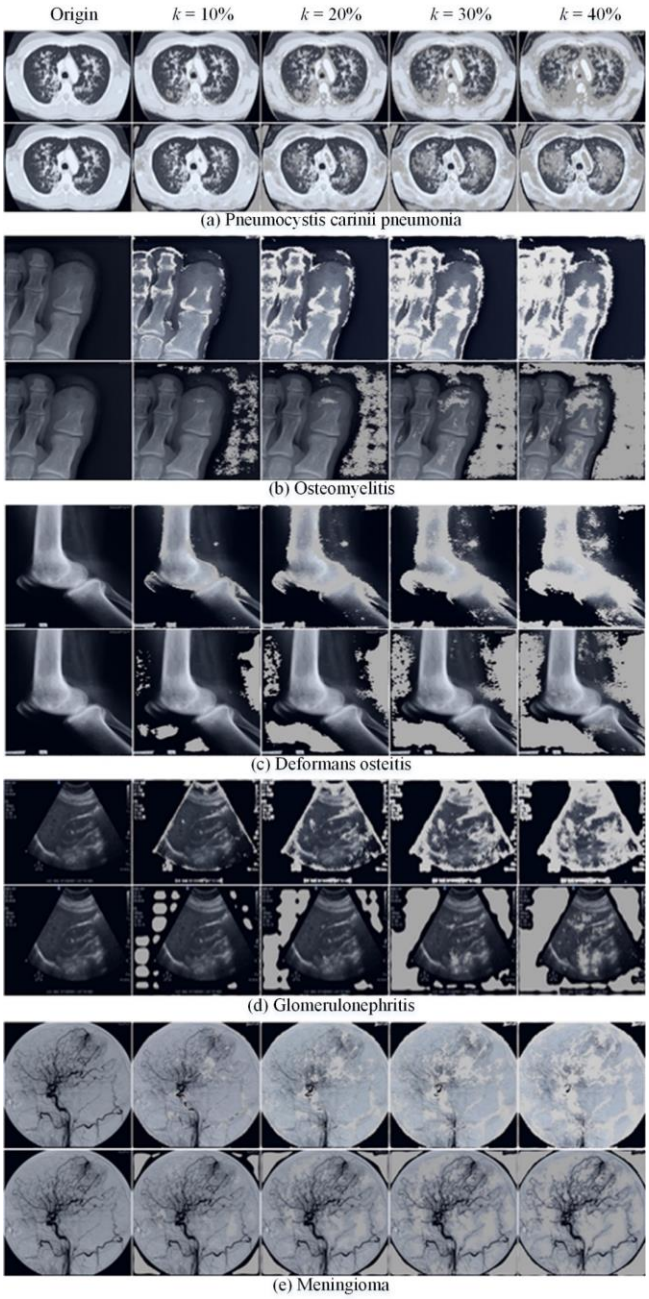


Fig. 5. Examples of counterfactual samples between causality and non-causality correlations.

Finally, Fig. 6 shows the causal saliency map generated by different visualization methods through model backpropagation

calculations. The data in VQA-Med-2020 are selected for this experiment. This data are invisible to the CCIS-MVQA model because the CCIS-MVQA model is only trained on VQA-RAD, SLAKE, and VQA-MED-2019 datasets.

From the perspective of human vision, the activation-based methods [52], e.g., Grad-CAM, Grad-CAM++, and Xgrad-CAM, illustrated in the second, third, and fourth columns in Fig. 6, use the linear combinations of class activation functions from convolutional layer to obtain causal saliency maps. These maps are excessively dispersed, blurry, noisy, and fail to visualize fine-grained features. However, these fine-grained features are significant in interpreting VQA-Med models [39].

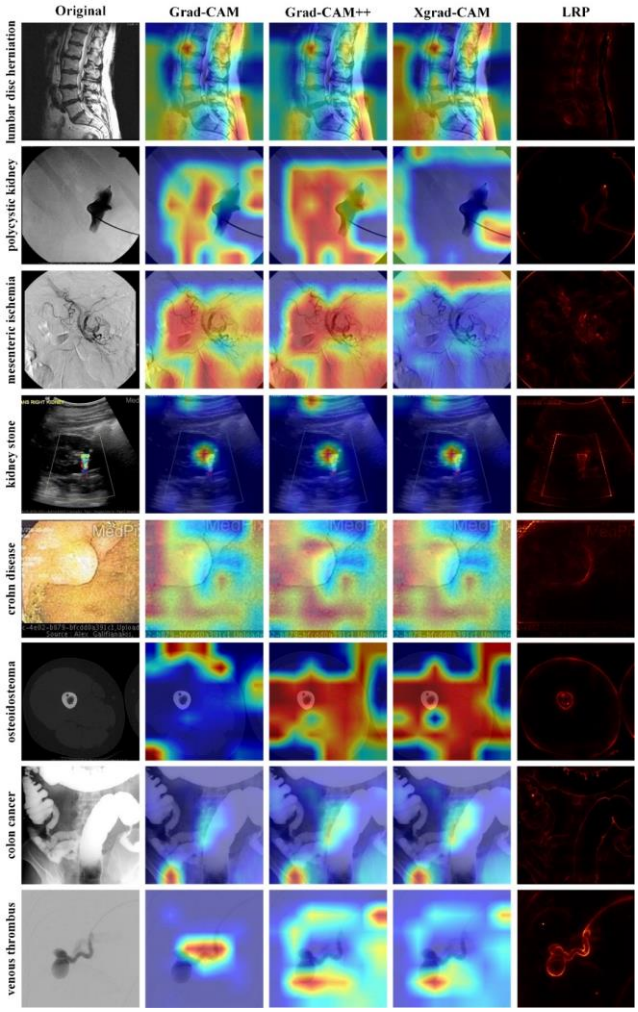


Fig.6. Examples for interpretable causal saliency map.

From the last column in Fig. 6, the LRP method does pay attention to the pixels with higher causal correlation. It generates a more precise causal saliency map for the original images with specific lesion or abnormality features, proving that the proposed CCIS-MVQA can obtain more accurate anomaly detection results. For example, the osteoid osteoma and venous thrombus are located more accurately in the sixth row, and the last row in Fig. 6. The abnormality or lesion is one of the major concerns during clinical practice, which becomes the key of VQA-Med. These examples in Fig. 6 show that the proposed CCIS-MVQA can capture the causal correlation between specific lesion or abnormality features in medical images and provide users with more appropriate explanations.

In addition, the activation-based approach neither guarantees that the interpretation is accurate nor reflects the model's decision-making process. LRP used in CCIS-MVQA calculates the contribution of each pixel of the input image by back-propagating the layer-wise relevance weights. Therefore, it is faithful to the model and highlights the target object rather than the background.

E. Quantitative evaluation of debiasing ability

We manually redevise the evenly distributed dataset VQA-RAD [48] to quantify the debiasing ability of the proposed CCIS-MVQA and construct a biased dataset VQA-RAD*. Specifically, for closed answers ("Yes/No"), the original distribution in VQA-RAD is 814 questions for an answer "No" and 899 questions for "Yes" in the training set. The test set had 133 questions for "No" and 118 for "Yes." The data distribution in VQA-RAD is relatively balanced. This paper divides the answers with similar original distribution into two groups roughly at a ratio of 1:3. Experimental results are shown in Table III.

According to Table III, the performances of all the models on VQA-RAD* are not as good as that of the original dataset. The performance of the MEVF series [25] decreases most obviously, and the accuracy of the opened, closed, and overall question answering decreases by 29%, 33%, and 32%, respectively. Compared with the two debiasing methods in generic domain VQA, i.e., RUBi [12] and CF-VQA [10], the accuracy of CCIS-MVQA's opened questions has declined similarly. In contrast, the overall performance on VQA-RAD* exceeds that of other models, indicating that our proposed CCIS-MVQA has a particular debiasing ability. The superscript values of our CCIS-MVQA are the MSE variance estimates for the accuracies.

TABLE III
THE QUANTITATIVE VERIFICATION RESULTS OF MODEL DEBIASING ABILITY (*ACC.* %)

Model	VQA-RAD			VQA-RAD*		
	Opened	Closed	Overall	Opened	Closed	Overall
MEVF-BAN[25] (2019)	43.90	75.10	62.60	31.14	50.62	42.83
MEVF-SAN[25] (2019)	40.70	74.10	60.80	32.21	50.47	43.17
RUBi [12] (2019)	63.46	78.22	69.83	51.53	64.27	59.17
CF-VQA [10] (2021)	60.26	74.08	68.55	51.95	**65.45**	60.05
CCIS-MVQA(ours)	**68.78**$^{\pm 0.23}$	**79.24**$^{\pm 0.16}$	**75.06**	**57.92**$^{\pm 0.31}$	61.68$^{\pm 0.27}$	**60.18**

*Represents a reconstructed dataset with opposite distributions of the training and test sets

Moreover, the performance degradation of the closed questions is higher than that of the open questions because the binary (Yes/No) answers of the closed questions accounts for a large proportion of VQA-RAD, and changing its distribution impacts the model performance. However, the answers to the opened questions are a relatively small proportion and a wide variety; changing the data distribution has relatively little effect on model performance.

RUBi [12] and CF-VQA [10] prevent the learning of question branches by influencing the model predictions, thereby dynamically adjusting loss functions to compensate for biases. We believe that the two methods eliminate the influence of the language modal to make up for language bias caused by the unbalanced dataset distribution and do not utilize the information of the image modal.

In addition, visualization technology [50] is used to demonstrate the model's testing process, as shown in Fig. 7. First, the answer distribution of two specific question modes is compared. Then, the extracted feature map is used to display the most critical area of the image in the test samples. MEVF [25] is used as the baseline model.

In the first row of Fig. 7, CCIS-MVQA shows its ability to suppress the unbalanced distribution for the question "*Is there an abnormality in the X-ray*?" which is a closed question with candidate answers of "Yes" or "No." The answer to most closed questions in the training set is "No." For test input, there is an abnormal bone density (red rectangle) in the shoulder joint, and the baseline model almost always answered "No" due to the unbalanced distribution. In contrast, CCIS-MVQA outputs an 80% probability of "Yes" and appears to infer the shoulder bone density abnormality by accurately locating the lesion region. In contrast, the baseline MEVF does not capture the abnormal region in the image and gets a wrong answer.

A similar result occurs in the second row in Fig. 7 with the question, "*What is an abnormality in the CT scan*?" Over 50% of answers in the dataset are "Cystic teratoma," and only 10% are "Colon cancer." For the test input, there is a tumor abnormality in the region of the colon (the red arrow points), and CCIS-MVQA accurately identifies the lesion. However, the baseline model MEVF captures the wrong location of the lesion. Regarding answer prediction, the baseline model only concludes with "Cystic teratoma" from the answer distribution in the training set. CCIS-MVQA applies a causal intervention strategy and deduces the correct "Colon cancer" answer based on the correct lesion region. However, the training samples of "colon cancer" are tiny.

The last row in Fig. 7 gives an example of the incorrect prediction of CCIS-MVQA, which could refer to the abnormal "small bowel volvulus" in the image. The dataset's uneven distribution affects the baseline model and answers "Hernia," which appears most frequently in the training set. Although CCIS-MVQA is not affected by the uneven distribution of datasets, it cannot distinguish the characteristics of "Small bowel lymphoma" and "Small bowel volvulus," which shows that the recognition ability of CCIS-MVQA to find features in images could still need to be further improved.

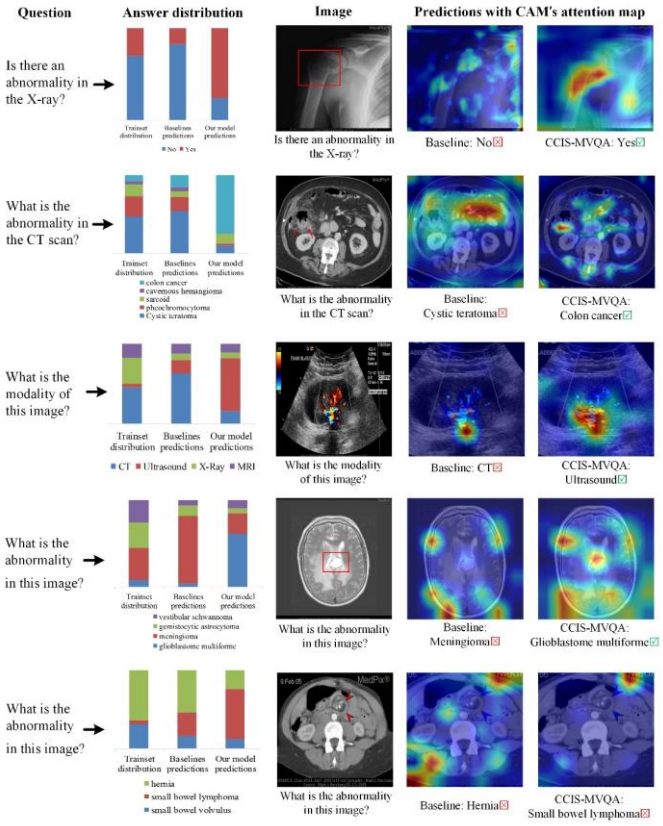


Fig.7. Visualization results of debiasing ability with CCIS-MVQA.

F. Ablation study and error analysis

We removed the CLIP module separately as our baseline model to evaluate the validity of the pre-training process. As shown in Table IV, the performance of the baseline model that removes the CLIP separately has decreased to varying degrees. Experimental results on VQA-Med-2019, SLAKE, and VQA-RAD datasets demonstrate the necessity and effectiveness of the pre-training process. The superscript value of our CCIS-MVQA model is the MSE of the generalization error.

TABLE IV
EFFECT OF PRE-TRAINING PROCESS ON CCIS-MVQA

CLIP	VQA-RAD (*Acc.* %)			SLAKE (*Acc.* %)		
	Opened	Closed	Overall	Opened	Closed	Overall
✓	**68.78**$^{\pm 0.23}$	**79.24**$^{\pm 0.16}$	**75.06**	**80.12**$^{\pm 0.11}$	**86.72**$^{\pm 0.07}$	**84.08**
×	49.86$^{\pm 0.19}$	76.79$^{\pm 0.21}$	66.08	76.61$^{\pm 0.18}$	80.16$^{\pm 0.12}$	78.84

CLIP	VQA-Med-2019				
	Modality	Plane	Organ	Abnormality	All
✓	**88.78**$^{\pm 0.38}$	**88.16**$^{\pm 0.21}$	**84.18**$^{\pm 0.14}$	**12.35**$^{\pm 0.40}$	**68.37**
×	81.39$^{\pm 0.64}$	79.13$^{\pm 0.29}$	77.86$^{\pm 0.22}$	4.16$^{\pm 0.54}$	60.60

We evaluated the effect of the batch size and the parameter k in the VQA-Med-2019 dataset. We set the batch size to 16, 32, 64, and 128, and set parameter k to 10, 20, 30, and 40%, respectively. Experimental results are illustrated in Fig. 8. We observe that the model has the best prediction performance when the batch size is 64. In addition, the prediction accuracy of the CCIS-MVQA model decreases gradually with the increase of parameter k.

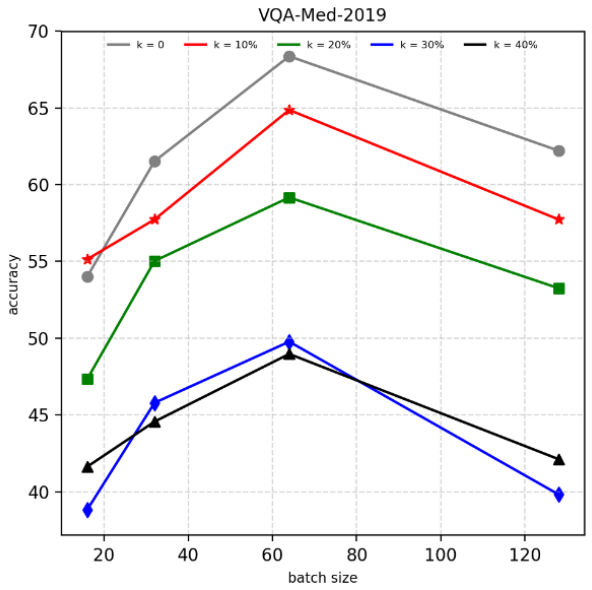


Fig. 8. Ablation studies on batch size and parameter *k*

Finally, we apply confusion metrics to evaluate the effect of the classifier model in the VQA-Med-2019 dataset and perform error analysis. We draw the confusion matrix for the Plane classifier with 16 candidate answers in Fig. 9.

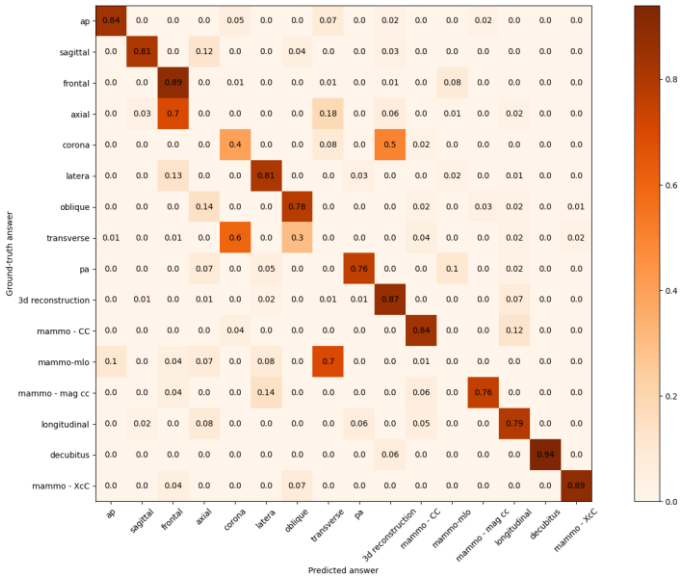


Fig.9. The confusion matrix of Plane

As shown in Fig. 9, most candidate answers are easily classified in the "plane" category, but some misclassifications exist. For example, the model tends to predict "axial" as "frontal" in that the two concepts are easily confused. Moreover, "transverse" and "mammo-mlo" cannot be accurately predicted, which indicates that the CCIS-MVQA model is not more robust in understanding images in lateral, internal, and external oblique positions.

V. Conclusions

This paper proposed a novel CCIS-MVQA framework to discuss how to mitigate the influence of language bias and enhance the interpretability of the VQA-Med model in mixed causal data. The proposed CCIS-MVQA framework consists of image feature extraction, question feature extraction, BAN, and interpretation generator. This paper focused on the interpretation generator to explore the interpretability and language bias of VQA-Med. We incorporated counterfactual interpretation and causal-effect reasoning into CCIS-MVQA to explore how the VQA-Med system responds to causal intervention strategy (such as the covered image of a given focus, how the model generates predictive answers), and quantify the effects of such intervention strategies.

The LRP technique was used in model training to generate counterfactual samples to obtain the causal connection between input samples and answers. Unlike other methods that use artificially specified rules, the counterfactual samples produced using LRP and the CCIS causal intervention strategy can generate interpretability and prediction results simultaneously. The causal intervention strategy follows a randomized controlled trial, and different conditions may produce different results.

Although our CCIS-MVQA model achieved good results in specific datasets, further improvement is needed to recognize fine abnormal features in the image. Future research could integrate existing medical knowledge bases and structured/unstructured knowledge to enhance performance. This could be achieved by incorporating knowledge graphs, large language models (LLMS) [53], and other knowledge bases into the training and inference process of VQA-Med models.

References

[1] A. Rajkomar, et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1-10, May 2018.

[2] P. Wang, T. Shi, and C. K. Reddy, "Text-to-SQL generation for question answering on electronic medical records," in *Proc. of WWW'20*, Taiwan, Apr. 2020, pp. 350-361.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, and D Parikh, "Vqa: Visual question answering," in *Proc.* IEEE/*CVF Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425-2433.

[4] M. Bansal, T. Gadgil, R. Shah, and P. Verma, "Medical Visual Question Answering at Image CLEF 2019-VQA Med.," in *Proc. Conf. and Labs of the Evaluation Forum* (*CLEF 2019*), Sept. 2019.

[5] M. Vu, R. Sznitman, T. Nyholm, and T. Löfstedt, "Ensemble of streamlined bilinear visual question answering models for the image clef 2019 challenge in the medical domain," in *Proc. Conf. and Labs of the Evaluation Forum* (*CLEF 2019*), Sept. 2019.

[6] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Trans. on Knowl. Discov. D.*, vol. 15, no. 5, pp. 1-46, May 2021.

[7] J. Pearl, *Causality: models, reasoning and inference*, 2nd Ed. New York: Cambridge University Press, 2009.

[8] J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, New York: Basic Books, 2018.

[9] D. Wang, Y. Yang, C. Tao, F. Kong, and L. Carin, "Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models," 2020, *arXiv:2012.03369*.

[10] Y. Niu, K. Tang, H. Zhang, Z. Lu, and J.R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12700-12710.

[11] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples

synthesizing for robust visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10800-10809.

[12]， R. Cadene, C. Dancette, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," Advances in neural information processing systems, vol. 32, 2019.

[13] R. R. Selvaraju, S. Lee，Y. Shen，H. Jin，and D. Parikh, "Taking a hint: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2591-2600.

[14] D. H. Park et al., "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8779-8788.

[15] X. Zhu, Z. Mao, C. Liu, P. Zhang, and Y. Zhang, "Overcoming language priors with self-supervised learning for visual question answering," 2020, arXiv:2012.11528,.

[16] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *Proc. IEEE Winter Conf. App. Comput. Vis.*, Jan. 2019, pp. 349-357.

[17] C. Fernandez, F. Provost, and X. Han, "Explaining data-driven decisions made by AI systems: the counterfactual approach," 2020, arXiv:2001.07417.

[18] D. Teney, E. Abbasnedjad, and A. V. D. Hengel, "Learning what makes a difference from counterfactual examples and gradient supervision," in *Proc. Euro. Conf. Comput. Vis.*, 2020, pp. 580-599.

[19] J. Pan, Y. Goyal, and S. Lee, "Question-conditioned counterfactual image generation for vqa," 2019, arXiv:1911.06352.

[20] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. EMNLP*, Nov. 2016, pp. 457-468.

[21] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21-29.

[22] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," Advances in neural information processing systems, vol. 31, 2018.

[23] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (*ICCV*), Oct. 2017, pp. 1821-1830.

[24] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947-5959, Dec. 2018.

[25] 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 Proc. *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* (*MICCAI*), Springer, 2019, pp. 522-530.

[26] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. (*MICCAI*), Cham, Switzerland: Springer, 2021, pp. 64-74.

[27] B. Liu, L. Zhan, and X. Wu, "Contrastive pre-training and representation distillation for medical visual question answering based on radiology images," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (*MICCAI*), Cham, Switzerland: Springer, 2021, pp. 210-220.

[28] S. Liu, X. Zhang, X. Zhou, and J. Yang, "BPI-MVQA: a bi-branch model for medical visual question answering," BMC Med. Imaging, vol. 22, no. 1, pp. 1-19, Apr. 2022.

[29] 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 *Proc. Int. Conf. Multimedia Retrieval*, Aug. 2021, pp. 456-460.

[30] F. Ren and Y. Zhou, "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," *IEEE Access*, vol. 8, pp. 50626-50636, Mar.2020.

[31] L. M. Zhan, B. Liu, L. Fan, J. Chen, and X. M. Wu, "Medical Visual Question Answering via Conditional Reasoning," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2345–2354

[32] M. H. Vu, T. Löfstedt, T. Nyholm, and R. Sznitman, "A Question-Centric Model for Visual Question Answering in Medical Imaging," *IEEE Trans. Med. Imaging*, vol. 39, no. 9, pp. 2856-2868, Sept. 2020.

[33] A. Zhang, W. Tao, Z. Li, H. Wang, and W. Zhang, "Type-Aware Medical Visual Question Answering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Apr. 2022, pp. 4838-4842.

[34] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, "AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering," *Knowl. Based Syst.*, vol. 255, Nov.2022, Art. no. 109763.

[35] Y. Yu, H. Li, H. Shi, L. Li, and J. Xiao, "Question-guided feature pyramid network for medical visual question answering," *Expert Syst. Appl.*, vol. 214, Mar.2023, Art. no. 119148.

[36] F. Cong, S. Xu, L. Guo, and Y. Tian, "Caption-Aware Medical VQA via Semantic Focusing and Progressive Cross-Modality Comprehension," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3569-3577.

[37] Y. Li et al., "A Bi-level representation learning model for medical visual question answering," *J. Biomed. Inform.*, vol. 134, Oct. 2022, Art. no. 104183.

[38] J. Huang et al., "Medical knowledge-based network for Patient-oriented Visual Question Answering," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103241.

[39] F. Cong, S. Xu, L. Guo, and Y. Tian, "Anomaly Matters: An Anomaly-Oriented Model for Medical Visual Question Answering," *IEEE Trans. Med. Imaging*, vol. 41, no. 11, pp. 3385-3397, Nov. 2022.

[40] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748-8763.

[41] S. Eslami, G. de Melo, and C. Meinel, "Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?" 2021, *arXiv:2112.13906*.

[42] T. He, Z. Zhang, H. Zhang, Z. Zhang, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 558-567.

[43] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in COntext (ROCO): a multimodal image dataset," *Lecture Notes in Computer Science*, Springer, 2018, pp. 180-189.

[44] J. Pearl, "Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution", in *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining* (*WSDM*), Feb. 2018.

[45] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, pp.1-46, Jul. 2015.

[46] L. Arras, G. Montavon, K. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," In *Proc. of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Sept. 2017, pp. 159-168.

[47] A. B. Abacha et al., "VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019," *Proc. Conf. and Labs of the Evaluation Forum* (*CLEF 2019*), Sept. 2019.

[48] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Sci. Data*, vol. 5, no. 1, pp. 1-10, 2018.

[49] B. Liu, et al., "Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *Proc. IEEE 18th Int. Symp. on Biomed. Imaging (ISBI)*, Apr. 2021, pp. 1650-1654.

[50] K. Gasmi, L. B. Ammar, K. Gasmi, G. Lejeune, H. Alshammari, and M.A. Mahmood, "Optimal deep neural network-based model for answering visual medical question," *Cybern. Syst.*, vol. 53, no. 5, pp. 403-424, Dec. 2022.

[51] S. Hooker, D. Erhan, P. J. Kindermans, and K. G. Brain, "A Benchmark for Interpretability Methods in Deep Neural Networks," in *Proc. NeurIPS*, Oct. 2019, pp.1-12.

[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (*ICCV*), Oct. 2017, pp. 618-62.

[53] C. Li, C. Wong, et al. "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," 2023, arXiv:2306.00890.