De-biasing multi-question learning in medical visual question answering via semantic-aware natural direct effect subtraction

1st Qishen Chen
Shanghai University
Shanghai, China
qs-chen@shu.edu.cn

2st Wenxuan He
Shanghai University
Shanghai, China
salieri@shu.edu.cn

3st Huahu Xu*
Shanghai University
Shanghai, China
huahuxu@shu.edu.cn

Abstract*—Medical Visual Question Answering (MedVQA) supports clinical decision-making by answering diagnostic questions based on medical images. Multi-Question Learning (MQL) has been adopted to jointly model multiple questions per image, improving performance by leveraging inter-question dependencies. However, MQL also amplifies language bias, where models rely on textual patterns rather than visual content, by introducing a new shortcut: reasoning from correlated questions alone. Using only questions, MQL can exhibit severe language bias, yet achieves 84.35% accuracy on the SLAKE dataset, nearly 30% higher than traditional single-question learning methods, highlighting the risk of spurious shortcuts in reasoning. Further experiments conducted on the proposed SLAKE-CP-Co dataset, designed to isolate the language bias introduced by co-occurring questions, provide evidence for the presence of multi-question language bias in the MQL. Thus, this paper proposes a novel causal framework for MedVQA under MQL. The proposed method explicitly identifies two types of language bias: the conventional Single-Question Language Bias and the newly introduced Multi-Question Language Bias. Both are mitigated by subtracting their Natural Direct Effects from the total effect, enforcing greater reliance on visual information. Additionally, this paper introduces a Semantic-Aware Bias Subtraction module that dynamically adjusts the subtraction strength based on question semantics, preserving beneficial language cues for domain-knowledge questions. Experiments on SLAKE, VQA-RAD, SLAKE-CP and SLAKE-CP-Co demonstrate that the proposed method outperforms state-of-the-art biased and de-biased models in both accuracy and robustness. The proposed approach offers a principled, effective strategy to balance multi-question reasoning and language bias mitigation, advancing the state of the art in MedVQA.

***Index Terms*—Language Bias, Medical Visual Question Answering, Multi-Question Learning**

I. Introduction

Medical visual question answering (MedVQA) aims to assist clinical decision-making by jointly reasoning over medical images and natural language questions. Despite recent progress, MedVQA systems remain vulnerable to language bias, where models exploit statistical regularities in questions and answers instead of relying on visual evidence. This issue is well documented in general VQA and is particularly problematic in the medical domain due to the high stakes of clinical interpretation and the subtle nature of medical imagery [1]–[3].

To improve diagnostic performance, recent MedVQA studies have adopted **Multi-question Learning (MQL)** [4], [5], where an image and all its associated questions are modeled jointly. By capturing correlations among related questions, MQL enables richer contextual reasoning and often outperforms traditional single-question learning. However, jointly modeling multiple questions also amplifies language bias: models may increasingly rely on cross-question co-occurrence patterns rather than visual content, leading to correct predictions through shortcut reasoning.

A closer examination of MedVQA datasets reveals that question co-occurrence can encode both beneficial semantic relations and harmful superficial correlations. Some question pairs naturally support each other through meaningful medical semantics, thereby facilitating accurate diagnosis, while others exhibit frequent but spurious correlations that do not require visual reasoning. As illustrated in Table I, the first type of question pair reflects a valid semantic dependency, where identifying the main organ enables subsequent reasoning about its function. In contrast, the second type (appears multiple 227 times in train set) shows a strong but superficial co-occurrence pattern in the training data (e.g., the frequent pairing of MRI and brain), even though image modality should not determine anatomical location. When such spurious correlations are exploited by MQL models, they introduce a new form of language bias that is absent in single-question settings. Existing de-biasing methods, which primarily suppress question-only shortcuts, are not designed to address this multi-question language bias.

Moreover, not all language bias in MedVQA is inherently undesirable. Medical datasets often include questions that can be answered correctly using general medical knowledge or common facts without reference to the image [6]. For such questions, linguistic priors are informative rather than misleading. However, most current de-biasing approaches indiscriminately suppress language information, which can degrade performance on knowledge-based questions. This highlights the need for a nuanced de-biasing strategy that selectively

TABLE I
TWO TYPES OF CO-OCCURRENCE QUESTION PAIR FROM SLAKE DATASET AND THEIR CORRESPONDING ANSWERS, THE FIRST PAIR CONTAINS BENEFICIAL SEMANTIC RELATIONS FOR ACCURATE DIAGNOSIS, WHILE MODELING THE RELATION SHIP OF THE SECOND PAIR LEADS TO SUPERFICIAL CLUE AND LANGUAGE BIASED REASONING.

No.	Question 1	Question 2
1	What is the largest organ in the picture? Lung	What is the function of the main organ in this picture? Breathe
2	What modality is used to take this image? MRI	Which part of the body does this image belong to? Brain

removes harmful shortcuts while preserving useful semantic cues.

To address these challenges, this paper proposes a causal de-biasing framework for multi-question MedVQA. By extending the conventional VQA causal model, we explicitly distinguish between two bias pathways: (1) single-question language bias and (2) multi-question language bias arising from question co-occurrence. Using causal effect decomposition, we derive a principled bias mitigation strategy that subtracts the direct influence of these bias pathways from model predictions. Furthermore, we introduce a semantic-aware mechanism that dynamically adjusts the degree of bias subtraction based on question content, allowing the model to retain legitimate linguistic information when appropriate. Extensive experiments on standard and bias-shifted MedVQA benchmarks demonstrate the effectiveness and robustness of the proposed approach.

The primary contributions of this paper are threefold.

- Causal modeling of multi-question bias: We propose a novel causal graph for MedVQA that explicitly models both single-question and multi-question language bias, providing the first formal treatment of bias amplification under MQL.
- Semantic-aware de-biasing mechanism: We introduce a content-adaptive bias subtraction strategy that selectively suppresses harmful language shortcuts while preserving useful semantic priors for knowledge-based questions.
- Experiments on multiple MedVQA benchmarks, including bias-shifted datasets, show that the proposed method consistently outperforms existing de-biasing approaches in accuracy and robustness.

II. RELATED WORK

Multi Question Learning in Medical Visual Question Answering: MedVQA is a multimodal task, where most research has focused on improving multimodal fusion between medical images and textual queries. For example, SAN [7] introduced a stacked attention mechanism that allows the model to query different image regions based on the input question. AMAM [8] proposed dual attention mechanisms, Image-to-Question and Word-to-Question attention, to capture fine-grained interactions between textual and visual modalities while simultaneously mining semantic information from the questions. More recently, Transformer-based architectures have gained popularity in MedVQA due to their ability to model intra-modal and cross-modal dependencies effectively [9]. However, these approaches commonly assume a one-to-one correspondence between image and question, overlooking the potential inter-question correlations that naturally occur when multiple queries are posed about the same image. In clinical scenarios, this assumption may not hold. Multiple co-occurring questions can provide additional context that improves diagnostic accuracy. For instance, questions indicating both the presence of bowel abnormalities and high air-fluid levels may jointly suggest Pneumoperitoneum. Similarly, identifying an image as belonging to the brain increases the likelihood of predicting brain tumor while reducing the likelihood of liver cancer. To capture such inter-dependencies, MQL was first proposed in general VideoQA tasks [5]. MQL jointly models multiple image-question pairs, generating a unified visual-question representation. In this approach, all textual and visual inputs for one image are compressed into a 1024-dimensional vector, enabling richer semantic representation across questions. MMQL [4] extended MQL to the MedVQA domain by reformulating the task as a masked language modeling problem. All questions for a given image are concatenated, and each answer is masked with a special token [MASK] for prediction, enabling the model to preserve individual question semantics while learning from collective context. This method has demonstrated strong performance as a baseline in MedVQA. In contrast, Rad-restruc [10] proposed a sequential reasoning framework in which questions are answered in hierarchical order. Basic-level questions (e.g., modality, plane, organ, presence) are predicted first, and their outputs are used to answer higher-order reasoning questions (e.g., size, location, attributes). However, this approach requires extensive additional annotations, and incorrect answers to early questions can propagate errors throughout the prediction pipeline.

Language Bias: Due to imbalanced data distributions, both general and medical VQA systems suffer from language bias, wherein the model learns to predict answers based solely on superficial language patterns, neglecting the visual modality. This can result in poor generalization, especially for rare or unseen samples. Fig. 1 shows the causal graph of traditional MedVQA methods, in which the dashed lines represent the inference shortcut, in other words, modality bias. To alleviate language bias, several methods have been proposed. Previous method using regularization technique in which a text-only branch estimates the dataset's language prior and subtracts it from the multimodal prediction [11]. RUBi [12], which employs a question-only model to generate a gating mask that reduces the influence of biased examples during training. Counterfactual VQA [13] framed the de-biasing problem using counterfactual inference, identifying the language bias as a direct causal effect and subtracting it from the total causal effect. In the MedVQA domain, similar efforts have emerged. DeBCF [2] and MedCFVQA [14] adapted counterfactual training to de-bias medical question answering, while CCIS-MVQA [15] constructed counterfactual samples by removing question-relevant regions from medical images. [16] treated

image-question pairs as positive samples and generated negative samples by mismatching either the image or the question, thereby disrupting spurious correlations. However, all of the above methods assume a **Single-question Learning (SQL)** setting, in which each question is answered independently. As shown in Fig. 2, introducing MQL amplifies language bias due to increased co-occurrence of related questions, potentially allowing the model to predict answers by exploiting question correlations alone. Therefore, new causal models and debiasing strategies are required to isolate and subtract the enhanced language priors introduced by MQL.

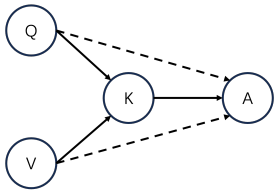

Fig. 1. The causal graph of vanilla MedVQA.

III. Methodology

A. Causal Graph

Let $D = \langle q_i, a_i, v_j \rangle$ denote the dataset consisting of n triples, where q_i, a_i, and v_j represent the question, answer, and image, respectively. A mapping function $g(\cdot)$ maps the question index $i \in \mathbb{N}^*, i < n$ to an image index $j \in \mathbb{N}^*, j < m$, where m is the total number of images. The MQL method adopted in this paper follows the approach proposed by MMQL. For each image, all associated questions are concatenated into a single long sentence to serve as the textual input. This formulation enables the model to capture both the semantic meaning of individual questions and the inter-question correlations. The process is defined in (1), where $\odot$ denotes the concatenation operation, and $f_t(\cdot)$ is the language encoder:

$$\begin{aligned} &\forall j < m, \exists Q_j,\ s.t.\ q_k \in Q_j,\ g(k) = j, \\ &T = q_1 \odot q_2 ... \odot q_k ... \odot q_{|Q_j|},\ q_k \in Q_j. \\ &\text{emb}_t = f_t(T) \end{aligned} \tag{1}$$

The resulting text embedding $\text{emb}_t = T_j$ is then fused with the visual embedding $\text{emb}v = f_v(v_j)$ to produce a multimodal representation emb_f. This fused embedding is used to predict the answers $a_i \in A_j$, where A_j denotes the set of answers corresponding to the question set $Q_j = q_1, q_2, \ldots, q_k, \ldots, q_{|Q_j|}$ for image v_j, and $g(k) = j$. Consequently, instead of modeling the conditional distribution $P(a_i \mid q_i, v_i)$ for each individual question-image pair, MQL models the joint distribution of $P(A_j \mid Q_j, v_j)$ which can be defined in (2):

$$\begin{aligned} K_1^{-k} &= Q_j \setminus \{q_k\}, \\ P(A_j \mid Q_j, v_j) &= \prod_{k=1}^{|Q_j|} P(a_k \mid Q_j, v_j), \\ &= \prod_{k=1}^{|Q_j|} P(a_k \mid q_k, K_1^{-k}, v_j), \end{aligned} \tag{2}$$

in which the K_1^{-k} is the correlation/context summary excluding the target question, which only carry cross-question information. Thus, to obtain the answer a_k, three factors are taken into consideration: target question q_k, cross-question information K_1^{-k}, and image v_j. Then, the (2) can be further written in the format of (3):

$$\begin{aligned} P(A_j \mid Q_j, v_j) = &\prod_{k=1}^{|Q_j|} [P(a_k \mid q_k) \cdot \frac{P(a_k \mid q_k, K_1^{-k})}{P(a_k \mid q_k)} \\ &\cdot \frac{P(a_k \mid q_k, K_1^{-k}, v_j)}{P(a_k \mid q_k, K_1^{-k})}] \\ = &\underbrace{\prod_{k=1}^{|Q_j|} P(a_k \mid q_k)}_{\text{SQLB}} \cdot \underbrace{\prod_{k=1}^{|Q_j|} \frac{P(a_k \mid q_k, K_1^{-k})}{P(a_k \mid q_k)}}_{\text{MQLB}} \\ &\cdot \underbrace{\prod_{k=1}^{|Q_j|} \frac{P(a_k \mid q_K, K_1^{-k}, v_j)}{P(a_k \mid q_k) \cdot P(a_k \mid q_k, K_1^{-k})}}_{\text{Image-grounded correction}}. \end{aligned} \tag{3}$$

The image-aware correction is normalized against the language-only bias. So, if the model prediction is independent of the image, then $P(a_k \mid q_j, K_1^{-k}, v_j) \approx P(a_k \mid q_k) \cdot P(a_k \mid q_k, K_1^{-k})$, image-aware correction is equal to 1. If the image provides strong corrective evidence, this ratio will deviate from 1, either boosting or reducing the influence of the language prior. Thus the correction term does not eliminate language bias, but rather measures deviation from it. Since the language bias is obtained, the causal graph can be construct as Fig. 2. In this paper $P(a_k \mid q_k, K_1^{-k})$ is considered as a new language bias: multi-question language bias ($K_1 \rightarrow A$ in Fig. 2), denoted as **Multi-question Language Bias (MQLB)** introduced by MQL method. Relatively, the vanilla single-question language bias ($Q \rightarrow A$ in Fig. 2) is denoted as **Single-question Language Bias (SQLB)**. In brief, the debiasing methods is aiming to removing the natural direct effect of node K_1 and Q.

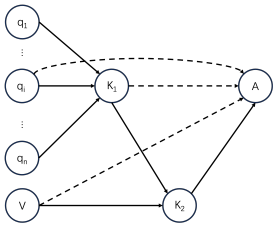

Fig. 2. The causal graph of MQL method for MedVQA.

B. De-bias Language Prior

To subtract the two types of language priors, this work follows the approach of Counterfactual VQA [13], which removes the Natural Direct Effect (NDE) from the Total Effect (TE). The resulting quantity is referred to as the Total Indirect

Effect (TIE), as defined in (4), and represents the causal effect with reduced language bias:

$$TIE = TE - NDE. \tag{4}$$

Based on the above definition, the MedVQA task with MQL can be represented using the causal graph illustrated in Fig. 2. In this graph, the answer set $A = a$ is influenced by the direct effects of the medical image $V = v$ and the set of questions $Q = \{q\}$. In addition, A is affected by the indirect effects introduced by the MQL process, denoted as $K_1(Q = \{q\})$, and the subsequent cross-modal fusion process $K_2(V = v, K_1 = k_1)$. Accordingly, the final prediction A is computed as described in (5). Notably, this paper adopts the RUBi approach [12], which has been shown to be a special case of the counterfactual framework in Counterfactual VQA, under the condition that the visual-to-answer path $V \rightarrow A$ (i.e., the NDE of V) is not considered. This simplification is motivated by the observation that visual bias is relatively negligible in the MedVQA domain.

$$A_{\{q\},k_1,k_2} = A(Q = \{q\},\ K_1 = k_1,\ K_2 = k_2). \tag{5}$$

Following the definition in Counterfactual VQA, the TE can be expressed as (6):

$$TE = A_{\{q\},k_1,k_2} - A_{\{q^*\},k_1^*,k_2^*}. \tag{6}$$

Here, $K_1^*(Q = \{q^*\})$ and $K_2^*(V = v^*,\ K_1 = k_1^*)$ represent the embeddings under the no-treatment condition, where q^* denotes the absence of the original question input. To subtract the language bias, NDE_{MQL} — representing the NDE of the MQLB — is defined as the effect when only the question set $\{q\}$ and the multi-question knowledge k_1 are given. The formal definition of NDE_{MQL} is provided in (7):

$$NDE_{MQL} = A_{\{q\},k_1,k_2^*} - A_{\{q^*\},k_1^*,k_2^*}. \tag{7}$$

Thus, based on (refeq6) and (7), the TIE of the MQLB can be expressed as shown in (8):

$$\begin{aligned} TIE_{MQL} &= TE - NDE_{MQL} \\ &= A_{\{q\},k_1,k_2} - A_{\{q\},k_1,k_2^*}. \end{aligned} \tag{8}$$

Similarly, the NDE of the SQLB can be defined when only the question set $\{q\}$ is given. In this paper, NDE_Q is used to represent the NDE of SQLB, and it can be calculated as shown in (9):

$$NDE_Q = \sum_{k=1}^{|Q_j|} A_{q_k,k_1^*,k_2^*} - A_{\{q^*\},k_1^*,k_2^*}. \tag{9}$$

Thus, based on the (6) and (9), the TIE of the SQLB can be expressed as shown in (10):

$$\begin{aligned} TIE_Q &= TE - NDE_Q \\ &= A_{\{q\},k_1,k_2} - \sum_{k=1}^{|Q_j|} A_{q_k,k_1^*,k_2^*}. \end{aligned} \tag{10}$$

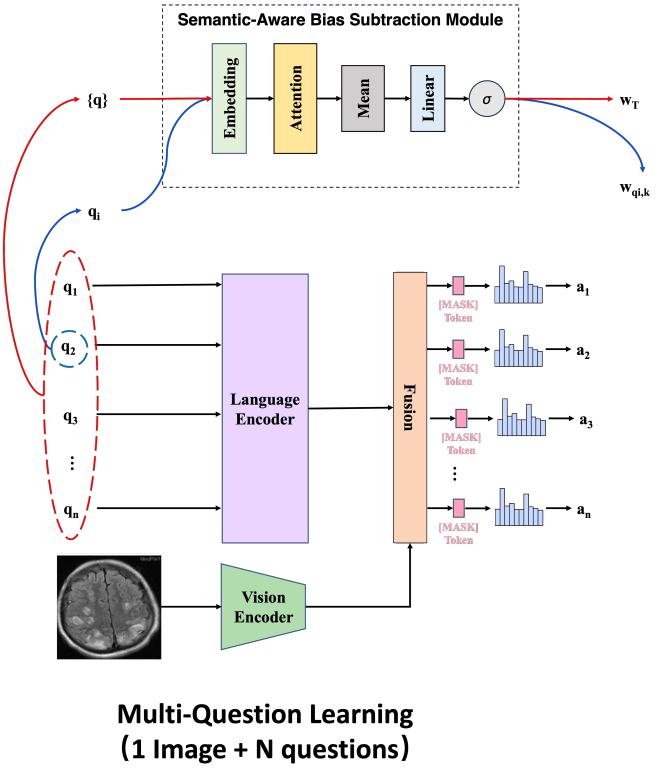

Fig. 3. The overall structure of the proposed model. The model is based on MMQL architecture, which takes one image and multiple related questions ($q_1 ... q_{|Q_j|}$), and output corresponding predictions ($a_1 ... a_{|Q_j|}$) in one forward pass. Each question has a [MASK] token (pink block) used for classification. In SABS module, q represents a set contains all questions and q_i represents single question, and SABS generate weights W_T for MQLB and $-w_{q_k}$ for SQLB.

Finally, the reasoning process is based on both TIE_Q and TIE_{MQL}. In this paper, the final prediction is obtained by summing these two TIEs, as defined in (11):

$$\begin{aligned} TIE &= TIE_{MQL} + TIE_Q \\ &= 2 * A_{\{q\},k_1,k_2} - A_{\{q\},k_1,k_2^*} - \sum_{k=1}^{|Q_j|} A_{q_k,k_1^*,k_2^*}. \end{aligned} \tag{11}$$

C. Semantic-Aware Bias Subtraction

Some questions query domain knowledge can, and should be answered solely based on the language modality. Therefore, to control the subtraction strength and preserve the useful language prior for such questions, this paper proposes a **Semantic-Aware Bias Subtraction Module (SABS)**. This module leverages a self-attention mechanism to extract the semantic meaning of each question and generates a weight for each question to control the degree of bias subtraction, which is illustrated in Fig. 3. Let the language encoder be defined as $f_t(\cdot)$. For a question q_k, the corresponding text embedding is denoted as $\text{emb}_{t,q_k} \in \mathbb{R}^{l \times d}$, where l is the maximum sequence length and d is the embedding dimension. The embedding emb_{t,q_k} is first processed through a self-attention layer, and its mean is then computed across the sequence dimension. A

linear layer $g^*(\cdot)$ subsequently maps the d-dimensional mean vector into a scalar weight parameter w_{q_k}. Finally, a sigmoid function is applied to normalize w_{q_k} into the range $(0, 1)$. The overall process of the Semantic-Aware Bias Subtraction Module is defined in (12):

$$\begin{aligned} \text{emb}_{t,q_k} &= f_t(q_{i_k}), \\ w_{q_k} &= g^*(\frac{\sum \text{emb}_{t,q_k}}{l}), \\ w_{q_k} &= \frac{1}{1+e^{-w_{q_k}}}. \end{aligned} \tag{12}$$

For MQLB, the single question q_k is replaced by the concatenation of questions T in (1). Thus, for MQLB the questions within the same medical image share the same weight w_T. Thus, the TIE is can be calculated by (13), with the introduction of the above weights. For the MQLB, the single question q_k is replaced by the concatenated question sequence T as defined in (1). Consequently, for MQLB, all questions associated with the same medical image share a common weight w_T. Incorporating this weight, the TIE is recalculated as shown in (13):

$$\begin{aligned} TIE =& 2 * A_{\{q\},k_1,k2} \\ & - w_T * A_{\{q\},k_1,k2^*} \\ & - \sum_{k=1}^{|Q_j|} w_{q_k} * A_{q_k,k_1^*,k2^*}. \end{aligned} \tag{13}$$

The SABS is designed to estimating the image-grounded correction term defined in Eq. (3). Specifically, the semantic-aware weights w_{q_k} and w_T modulate the subtraction of language-only effects, while the image-grounded correction $P(a_k \mid q_k, K_1^{-k}, v_j)/P(a_k \mid q_k, K_1^{-k})$ is preserved to capture deviations introduced by visual evidence. In this way, bias subtraction does not directly suppress language predictions, but regulates their contribution relative to the image-conditioned reasoning. This coupling ensures that harmful language shortcuts are attenuated, while visual corrections remain effective when image information is required.

IV. Experiments Results

A. Implementation Detail

Backbone: In Subsection III-B, v^* and $\{q^*\}$ are defined as the conditions where no image or question is provided. However, neural networks cannot directly process void inputs. To address this, this paper assumes that under such conditions, the model performs random guessing with equal probability across possible answers. As a result, the void input can be replaced by a constant c that induces uniform output probabilities. The value of c is estimated by minimizing the KL divergence. Then, this paper follows the backbone design of MMQL [4]. The visual encoder used is Swin-Transformer-small [17], and the language encoder is BERT-base-uncased [18]. The visual and textual embeddings are fused using a six-layer Transformer architecture, comprising three encoder layers and three decoder layers. Conditional reasoning [19] is employed after the fusion stage, providing independent classifier heads for open-ended and close-ended questions. Additionally, an independent BERT-base-uncased model is used as the language encoder for the proposed Semantic-Aware Bias Subtraction module. The overall method proposed in this paper is referred to as **De-bias-MMQL**.

Hyper-parameter: In the experiments, the maximum sequence length is set to 430 for the VQA-RAD dataset and 300 for the SLAKE (and SLAKE-CP) datasets, corresponding to a maximum of 22 and 20 questions per image, respectively. AdamW is used as the optimizer. The learning rate is set to $2.5e-5$ for the language encoders and $5e-5$ for the rest of the model. The batch size is set to 8. The two weighting parameters. The two weighting parameters, λ_1 and λ_2, used in the classification loss L_{cls}, are both set to 0.3. All experiments are conducted on an Ubuntu 22.04 server equipped with a single A100 GPU (40 GB).

B. Datasets and Evaluation Metrics

VQA-RAD: VQA-RAD [20] is a radiology-focused dataset manually annotated by individuals with professional medical expertise. It contains 315 radiological images covering three anatomical regions—head, chest, and abdomen—accompanied by a total of 3,515 clinical questions, averaging approximately 10 questions per image.

SLAKE: SLAKE [6] is a bilingual (Chinese and English) visual question answering dataset, offering broader organ system coverage compared to VQA-RAD. It comprises 642 medical images and over 14,000 questions, including both open-ended and close-ended formats. To maintain consistency with previous studies, this paper focuses on the English subset, which contains approximately 7,000 questions, averaging 11 questions per image. The SLAKE dataset is split at the image level into training (70%, 450 images), validation (15%, 96 images), and testing (15%, 96 images) sets.

SLAKE-CP: Following the work of DeBCF [2], a bias-sensitive MedVQA dataset named SLAKE-CP is constructed by re-splitting the original SLAKE dataset.

SLAKE-CP-Co: SLAKE-CP-Co aims to measure the MQLB in multi-question learning methods, by changing the answer distribution of commonly appeared question pairs. In this paper, SLAKE-CP-Co is generated under the following procedure:

- The process begin by parsing the original SLAKE training split and grouping all question–answer entries by their corresponding image. For each image, this paper enumerate all possible pairs of questions associated with that image. These pairs are recorded using a frequency counter to determine how often each question pair appears across the dataset. To avoid redundancy, question pairs are treated as unordered, i.e., (q_1, q_2) is considered the same as (q_2, q_1). After processing the entire training set, top 50 most frequently co-occurring question pairs are selected.
- After identifying the top 50 most frequent question pairs, this paper further analyze the joint answer distribution of each pair to identify cases with strong co-occurrence

patterns. For each question pair (q_1, q_2), this paper compute the empirical joint distribution of their corresponding answers across all instances where both questions appear with the same image. The distribution is represented as a mapping from answer pairs (a_1, a_2) to their occurrence counts and normalized probabilities.

- Then, the top co-occurring question pairs are further filtered according to two strict criteria: (1) The two questions in each pair must be semantically independent. That is, the answer to one question should not logically imply or influence the answer to the other. This guarantees that any observed correlation in their answer distributions is not due to intrinsic causal relationships. (2) The answer co-occurrence patterns must be significantly biased—specifically, exhibiting rare or near-absent combinations of certain answers.
- After applying these criteria, a new dataset is synthesize by selectively constructing samples with rare answer combinations that are either underrepresented or non-existent in the original dataset. For each image in the test set, the synthesize process filters for cases that satisfy the question pair and their answers. These samples are added to a new dataset. Finally at image level duplicated questions are removed.

In this paper totally 458 samples are generated with 175 open-ended questions and 283 close-ended questions.

C. Compare to Existing Methods

Table II presents the comparison results of existing methods on the VQA-RAD and SLAKE datasets. To ensure consistency with previous works, Table II reports the accuracy for open-ended and close-ended questions separately, as well as the overall accuracy regardless of question types.

TABLE II
COMPARISON OF DIFFERENT MED-VQA MODELS ON THE VQA-RAD AND SLAKE DATASETS. THE COMPARED METHODS ARE CATEGORIZED INTO TWO GROUPS: BIASED METHODS (UPPER SECTION) AND DE-BIASED METHODS (LOWER SECTION). NUMBERS IN PARENTHESES INDICATE THE IMPROVEMENT RELATIVE TO THE SECOND-BEST MODEL IN EACH COLUMN.

Models	VQA-RAD			SLAKE		
	Open	Close	Overall	Open	Close	Overall
MQL∗	39.66	74.26	60.53	73.64	69.23	71.91
MEVF-BAN-CR	55.87	80.88	70.95	78.76	81.97	80.02
MMBERT	63.13	77.94	72.06	-	-	-
CMSA	61.45	80.88	73.17	73.80	81.97	77.00
RAMM	63.69	79.80	73.39	-	-	-
M-Mixup	55.31	79.04	69.62	80.16	86.06	82.47
PubMedCLIP	58.66	80.88	72.06	77.83	81.49	79.26
LGPFN	78.17	81.57	80.39	81.09	88.70	84.07
MMQL	64.25	85.66	77.16	84.19	90.63	86.71
DeBCF	58.66	80.88	72.06	80.78	84.86	82.38
MedCFVQA	48.60	61.40	56.32	82.02	89.90	85.11
Tri-VQA	60.34	82.72	73.84	81.55	85.58	83.13
CNSP	68.16	85.66	78.71	82.79	86.54	84.26
CCIS-MVQA	68.78	79.24	75.06	80.12	86.72	84.08
De-bias-MMQL (Ours)	**73.18** (-4.99)	**87.13** (+1.47)	**81.60** (+1.21)	**86.05** (+1.86)	**91.59** (+0.96)	**88.22** (+1.51)

Note: * stands for re-implementation

The MMQL method achieves an overall accuracy of 77.16% on VQA-RAD and 86.71% on SLAKE, representing the highest score among biased methods. It outperforms strong baselines such as MEVF-BAN-CR [19], CMSA [21] and LGPFN [22], pre-training-based methods like MMBERT [9] and PubMedCLIP [23], retrieval-augmented methods such as RAMM [24], and data augmentation based methods like M-Mixup [25]. The strong performance of MMQL highlights the potential of multi-question learning as a simple yet effective baseline for future MedVQA research. However, the relatively low performance of the original MQL [5] method for video understanding tasks indicates the necessity of adapting MQL approaches specifically for the medical domain. Furthermore, MMQL surpasses most de-biasing methods (except CNSP [16] on VQA-RAD), largely because existing de-biasing methods often use less powerful backbones. This suggests the importance of adopting stronger base models when developing de-biasing strategies. The high performance of CNSP likely benefits from the use of ViT and BERT architectures, similar to those employed by MMQL, and the introduction of counterfactual negative samples during pre-training, which is particularly effective for small datasets like VQA-RAD.

Compared to MMQL, the proposed De-bias-MMQL further improves diagnostic accuracy by 2.66% on VQA-RAD and 1.51% on SLAKE. For DeBCF, MedCFVQA, and Tri-VQA [1], these methods operate under the SQL paradigm, where only the SQLB exists during inference. By introducing MQL and the Semantic-Aware Bias Subtraction strategy to simultaneously address both MQLB and SQLB, De-bias-MMQL leverages the advantages of multi-question reasoning while maintaining low language bias, thus outperforming the aforementioned methods. Moreover, compared to CNSP and CCIS-MVQA, the proposed De-bias-MMQL achieves higher accuracy without the need for generating negative samples, resulting in a shorter and more efficient training process.

Furthermore, the proposed De-bias-MMQL is compared with existing de-biasing methods on the changing priors dataset, SLAKE-CP. The training and test sets of SLAKE-CP are constructed with different prior distributions of answers, aiming to encourage models to focus on interpretable visual features rather than memorizing dataset biases. The results are presented in Table III.

TABLE III
COMPARISON OF DIFFERENT DE-BIAS MED-VQA MODELS ON SLAKE-CP.

Models	SLAKE-CP		
	Open	Close	Overall
DeBCF	18.76	34.38	24.88
MedCFVQA	18.45	33.17	24.22
Tri-VQA	21.24	37.74	26.58
CNSP	19.84	39.66	27.62
CCIS-MVQA	21.71	41.35	29.41
De-bias-MMQL (Ours)	**23.10** (+1.39)	**43.99** (+2.64)	**31.29** (+1.88)

When the distribution inconsistency between the training and test sets is enlarged, all methods experience significant performance drops. Methods based on negative sample construction, such as CNSP and CCIS-MVQA, demonstrate

higher robustness under distribution shifts, achieving overall accuracies of 27.62% and 29.41%, respectively. The use of negative samples helps guide models to focus on relevant regions that contribute most to diagnostic reasoning. On the other hand, causal inference-based methods, which aim to subtract language bias, are generally less effective under distribution shifts, with overall accuracies falling below 27%. Notably, although DeBCF incorporates both causal inference and negative sample construction, it still records the lowest accuracy among all compared methods. This performance gap may be attributed to its adoption of less powerful encoders (ResNet and LSTM) rather than more advanced architectures like ViT and BERT. The proposed De-bias-MMQL achieves a 1.88% improvement compared to the second-best model. This gain may result from the integration of multi-question learning, which allows the model to better capture the co-occurrence patterns among multiple questions, even when the distribution of individual questions is altered.

D. Ablation Study

To verify the effectiveness of the proposed modules, an ablation study is conducted on the SLAKE dataset. The experiments are structured into five stages. First, the overall accuracy of the MMQL model is reported as the baseline. Second, the single-question language bias and the multi-question language bias are independently removed. Third, both single-question language bias and the multi-question language bias are jointly removed from the MMQL baseline. Finally, based on the previous model, the Semantic-Aware Bias Subtraction module is introduced.

TABLE IV
ABLATION STUDY ON SLAKE. SQLB DENOTES THE REMOVAL OF SINGLE-QUESTION LANGUAGE BIAS, MQLB DENOTES THE REMOVAL OF MULTI-QUESTION LANGUAGE BIAS, AND SABA REFERS TO THE PROPOSED SEMANTIC-AWARE BIAS SUBTRACTION MODULE.

Base Model	SQLB	MQLB	SABS	Overall Accuracy
MMQL				86.71
	✓			85.96
		✓		86.99
	✓	✓		87.84
	✓	✓	✓	88.22

The results are summarized in Table IV. Removing only the SQLB results in a 1.75% decrease in accuracy compared to the baseline, indicating that simply eliminating SQLB does not improve performance. In contrast, removing only the MQLB yields a marginal improvement of 0.28%. However, removing both SQLB and MQLB together leads to a 1.13% improvement over the baseline, demonstrating that addressing both types of biases simultaneously is necessary for effective de-biasing. These findings suggest that language biases exist in both shortcut paths $Q \rightarrow A$ and $K_1 \rightarrow A$ as shown in Fig. 2, and that optimal performance requires the removal of both. Finally, introducing the Semantic-Aware Bias Subtraction module provides an additional 0.38% performance gain. Since the bias subtraction weight is constrained between 0 and 1, this improvement supports the hypothesis that some domain knowledge questions can be answered solely based on textual information, and thus benefit from preserving part of the language bias. Overall, these results validate the effectiveness of the proposed causal model and the Semantic-Aware Bias Subtraction module.

TABLE V
ABLATION STUDY ON SLAKE-CP-CO. SQLB DENOTES THE REMOVAL OF SINGLE-QUESTION LANGUAGE BIAS, MQLB DENOTES THE REMOVAL OF MULTI-QUESTION LANGUAGE BIAS, AND SABA REFERS TO THE PROPOSED SEMANTIC-AWARE BIAS SUBTRACTION MODULE.

Base Model	SQLB	MQLB	SABS	Open	Close	Overall
MMQL				32.57%	66.07%	53.28%
	✓			32.57%	63.96%	46.63%
		✓		38.85%	70.67%	58.51%
	✓	✓		35.43%	69.61%	56.55%
	✓	✓	✓	37.71%	69.61%	57.42%

Table V presents an ablation study on the counter-prior dataset SLAKE-CP-Co, assessing the effectiveness of various bias removal strategies. The baseline model is MMQL, which is evaluated alongside different combinations of SQLB, MQLB removal, and the proposed SABS module. The removal of MQLB alone yields the most significant improvement, increasing open-ended accuracy by 6.3%, close-ended accuracy by 4.6%, and overall accuracy by 5.2% compared to the baseline. This is expected, as SLAKE-CP-Co was explicitly constructed to reverse answer priors between co-occurring question pairs. Thus, eliminating the multi-question shortcut compels the model to rely on visual evidence, resulting in a marked performance gain. In contrast, removing only SQLB leads to a performance drop: close-ended accuracy decreases by 2.1%. This suggests that, in the counter-prior split, some single-question priors still encode useful domain knowledge or reflect natural class distributions. Removing these helpful cues indiscriminately negatively impacts model predictions. When both SQLB and MQLB are removed, the model improves over the baseline (+3.3% overall), but underperforms compared to the MQLB-only setting. This indicates that the benefit of removing MQLB is partially offset by the harm caused by discarding beneficial SQLB. Integrating SABS after joint SQLB and MQLB subtraction recovers some performance. Open-ended accuracy increases by 2.3%, and overall accuracy improves by 0.9% compared to the dual-bias removal setup. This demonstrates that SABS can reintroduce valid question-dependent priors in a controlled, semantically-aware manner. Nevertheless, this configuration still does not outperform the simpler MQLB-only removal, indicating room for improvement in SABS's weighting strategy. The results highlight the importance of targeted de-biasing. Specifically, removing harmful multi-question priors is crucial for counter-prior robustness. However, blindly removing all linguistic cues can degrade performance, underscoring the need for selective and semantics-aware strategies such as SABS to retain beneficial priors while eliminating harmful correlations.

V. Conclusion

In this paper, a novel causal model is proposed for multi-question learning in the MedVQA domain. Compared to traditional causal models, a new shortcut $Q \rightarrow K_1 \rightarrow A$ is introduced to represent the language bias induced by the co-occurrence of multiple questions under a single image. Based on the proposed causal graph, the language biases arising from both single-question and multi-question interactions are mitigated by subtracting the natural direct effect from the total effect. Furthermore, to accommodate domain knowledge questions that can be answered solely based on textual information, a Semantic-Aware Bias Subtraction module is proposed to dynamically adjust the strength of the bias subtraction. Experiment results shows the proposed method surpasses both biased and de-biased methods under both traditional distributions and changing prior distributions.

References

[1] L. Fan, X. Gong, C. Zheng, and Y. Ou, "Tri-vqa: Triangular reasoning medical visual question answering for multi-attribute analysis," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 1485–1488.

[2] C. Zhan, P. Peng, H. Zhang, H. Sun, C. Shang, T. Chen, H. Wang, G. Wang, and H. Wang, "Debiasing medical visual question answering via counterfactual training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 382–393.

[3] M. U. Shoukat, L. Yan, J. Zhang, Y. Cheng, M. U. Raza, and A. Niaz, "Smart home for enhanced healthcare: exploring human machine interface oriented digital twin model," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 31 297–31 315, 2024.

[4] Q. Chen, M. Bian, and H. Xu, "Mmql: Multi-question learning for medical visual question answering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 480–489.

[5] C. Lei, L. Wu, D. Liu, Z. Li, G. Wang, H. Tang, and H. Li, "Multi-question learning for visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 328–11 335.

[6] 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)*. IEEE, 2021, pp. 1650–1654.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[8] 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," *Knowledge-Based Systems*, vol. 255, p. 109763, 2022.

[9] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "Mmbert: Multimodal bert pretraining for improved medical vqa," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1033–1036.

[10] C. Pellegrini, M. Keicher, E. Özsoy, and N. Navab, "Rad-restruct: A novel vqa benchmark and method for structured radiology reporting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 409–419.

[11] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," *Advances in neural information processing systems*, vol. 31, 2018.

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

[13] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 700–12 710.

[14] S. Ye, U. Naseem, M. Meng, D. Feng, and J. Kim, "A causal approach to mitigate modality preference bias in medical visual question answering," in *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, 2024, pp. 13–17.

[15] L. Cai, H. Fang, N. Xu, and B. Ren, "Counterfactual causal-effect intervention for interpretable medical visual question answering," *IEEE Transactions on Medical Imaging*, 2024.

[16] H. Cong, L. Liu, X. Yang, W. Peng, and L. Liu, "Constructing negative sample pairs to mitigate language priors in medical visual question answering," in *Proceedings of the 2024 5th International Conference on Intelligent Medicine and Health*, 2024, pp. 121–127.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[19] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2345–2354.

[20] 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*, vol. 5, no. 1, pp. 1–10, 2018.

[21] 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*, 2021, pp. 456–460.

[22] S. Du, S. Liang, and Y. Gu, "A language-guided progressive fusion network with semantic density alignment for medical visual question answering," *Journal of Biomedical Informatics*, vol. 165, p. 104811, 2025.

[23] 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*, 2023, pp. 1151–1163.

[24] Z. Yuan, Q. Jin, C. Tan, Z. Zhao, H. Yuan, F. Huang, and S. Huang, "Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training," *arXiv e-prints*, pp. arXiv–2303, 2023.

[25] 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)*. IEEE, 2022, pp. 2648–2655.