

A UNIFIED MULTI-DATASET FRAMEWORK FOR MEDICAL VISUAL QUESTION ANSWERING VIA PRETRAINED TRANSFORMERS AND CONTRASTIVE LEARNING

Bao-Nguyen Quoc, Huy-Ho Huu, Khang-Nguyen Hoang Duy
and Thu-Le Vo Minh

FPT University, Ho Chi Minh Campus, Vietnam

ABSTRACT

Medical Visual Question Answering (Med-VQA) aims to generate accurate answers to clinical questions grounded in medical images. However, existing models often struggle with limited generalization across datasets and insufficient understanding of specialized medical terminology. In this work, we propose a unified multi-dataset Med-VQA framework that integrates general-purpose vision-language models (e.g., BLIP) with domain-specific language models such as BioGPT to better capture biomedical semantics. Our architecture introduces a novel Mixture-of-Experts (Med-MoE) module that fuses knowledge across modalities and datasets, and it is jointly optimized using contrastive loss, image-text matching, and language modeling objectives. By combining cross-dataset supervision with domain-aware components, our approach achieves improved reasoning and generalization. Experimental results on VQA-RAD and PathVQA demonstrate state-of-the-art performance, validating the effectiveness of our unified framework.

KEYWORDS

Vision Transformer (ViT), Medical VQA, Transformer, PathVQA, VQA-RAD

1. INTRODUCTION

Medical Visual Question Answering (VQA) systems hold tremendous potential for supporting clinical decision-making by automatically interpreting medical images and answering clinically relevant questions. However, this domain presents unique challenges that general VQA approaches fail to address effectively:

- The complexity and specificity of medical terminology
- The high variability across medical imaging modalities (pathology, radiology, etc.)
- Limited availability of annotated medical data
- The critical need for domain-specific knowledge and reasoning

In this paper, we propose a unified multi-dataset framework for medical VQA that overcomes these challenges through several key innovations:

- **Novel Multimodal Architecture Integration:** We introduce a unified framework that synergistically combines BLIP’s vision transformer for robust medical image understanding, BERT’s contextual language processing for clinical questions, and BioGPT’s domain-specialized language generation—a combination specifically engineered to handle the complexity of medical VQA.
- **Medical-Specific Mixture-of-Experts (Med-MoE):** We present a specialized MoE module designed specifically for medical visual-textual reasoning, with dynamic routing optimized for clinical context preservation.
- **Multi-Objective Training Strategy:** We develop a comprehensive training methodology that combines contrastive learning, image-text matching, and autoregressive language modeling to enforce robust cross-modal alignment specifically for medical imaging and terminology. **Cross-Domain Medical Knowledge Transfer:** Our approach leverages transfer learning across distinct medical imaging domains (pathology and radiology), enabling the model to build a unified representation of medical visual-textual knowledge.
- **Custom Vocabulary Construction:** We introduce a medical terminology-focused vocabulary construction method that enhances the representation of clinical terms, improving text processing and understanding.

Our approach differs from previous medical VQA systems in its end-to-end design that processes multiple medical imaging modalities while handling diverse question-answer formats within a single model architecture. Through extensive evaluation on the PathVQA and VQA-RAD datasets, we demonstrate competitive performance and provide insights into model behavior across different question types and imaging modalities.

2. RELATED WORK

Visual Question Answering (VQA) has rapidly evolved from early methods based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to more advanced transformer-based architectures. Early VQA systems primarily combined CNNs for visual feature extraction with RNNs for language processing. However, these approaches were limited by their inability to model long-range dependencies and complex interactions between modalities.

Recent advances in transformer-based models, such as the Vision Transformer (ViT) [3] and large-scale multimodal pretraining frameworks like CLIP [18] and BLIP [2], have revolutionized VQA by enabling more effective joint modeling of vision and language. CLIP demonstrated that learning from massive amounts of image–text pairs can produce powerful cross-modal representations, while BLIP further improved on these methods by incorporating a bidirectional language model and a more sophisticated image encoder. These works have set a new benchmark for general-domain VQA, but their direct application to the medical domain remains challenging due to domain-specific language and visual nuances.

In the medical domain, specialized datasets such as PathVQA [6] and VQA-RAD [7] have been introduced to address the unique challenges of clinical image interpretation and question answering. The PathVQA dataset focuses on pathology images and provides extensive annotations that cover both open-ended and closed-ended questions. VQA-RAD, on the other hand, offers a benchmark for radiology by presenting clinically relevant questions over a limited but high-quality set of radiological images. These datasets have spurred the development of methods that specifically target the intricacies of medical VQA.

Several recent studies have proposed innovative approaches for medical VQA. For instance, MEVF [13] addresses the data scarcity problem by leveraging data augmentation and domain adaptation techniques. MMQ [14] and VQAMix [15] propose meta-model and mixup-based strategies, respectively, to enhance the robustness of VQA models against the heterogeneous nature of medical images and language. AMAM [16] introduces an asymmetric cross-modal attention mechanism with multimodal augmented mixup to better capture the nuanced relationships between medical images and questions. In addition, models such as M3AE [21] and MUMC [22] further refine cross-modal alignment by integrating contrastive learning and advanced autoregressive language modeling. These works utilize various loss functions, such as image-text matching (ITM) loss and contrastive loss, to improve the fusion of visual and textual representations. Furthermore, BioGPT [5] and related methods highlight the importance of domain-specific language models in generating clinically accurate and coherent responses.

Despite these advances, many existing methods either rely on single-dataset training or lack the ability to generalize across different medical imaging modalities. Moreover, while some approaches have explored the incorporation of dynamic fusion mechanisms such as Mixture-of-Experts (MoE), their benefits in the medical VQA context have been limited. Our work addresses these gaps by proposing a unified multi-dataset framework that integrates the strengths of state-of-the-art pretrained models (BLIP, BERT, and BioGPT) and employs a robust training strategy combining contrastive, ITM, and autoregressive language modeling losses. In doing so, our model not only achieves competitive performance on both open-ended and closed-ended questions but also demonstrates improved generalizability across the diverse clinical settings represented in the PathVQA and VQA-RAD datasets.

By leveraging multi-dataset pretraining, domain-specific text processing with a custom vocabulary, and carefully designed loss functions, our approach advances the field of medical VQA. It provides a comprehensive solution that bridges the gap between the general success of transformer-based VQA systems and the unique challenges posed by medical imaging and clinical language. Future work will explore further refinements in multimodal fusion and fine-tuning strategies to address the remaining performance gaps, particularly in the closed-ended task.

Table 1. Comparison of our approach with previous medical VQA method.

Method	Vision Encoder	Text Encoder	Fusion Mechanism	Loss Functions
MEVF [13]	ResNet-152	LSTM	Attention	CE
MMQ [14]	ResNet-152	BiLSTM	Concatenation	CE
VQAMix [15]	DenseNet-121	LSTM	MLP	CE + CL
AMAM [16]	ResNet-101	LSTM	Cross-modal attention	CE
M3AE [21]	ViT-B/16	BERT	Cross-attention	CE + CL + MLM
MUMC [22]	Swin-B	BERT	Co-attention	CE + CL
Ours	BLIP (ViT-B/16)	BERT + BioGPT	Med-MoE + Cross-attention	CE + CL + ITM + LM

This comparison highlights the unique combination of components in our approach. Unlike previous methods, we leverage BLIP's strong visual foundation while incorporating domain-specialized language models (BioGPT) and a medical-adapted Mixture-of-Experts fusion mechanism. Our approach also employs a more comprehensive set of training objectives and domain adaptation techniques, specifically designed to address the challenges of medical VQA.

3. DATASETS

In this study, two benchmark datasets are utilized: PathVQA, introduced by He et al. in 2020 [6], and VQA-RAD, presented by Lau et al. in 2018 [7]. Both datasets focus on the medical domain but differ in terms of image modalities, annotation scope, and question types.

3.1. PathVQA Dataset

The PathVQA dataset, introduced by He et al. [6], is the first large-scale VQA dataset focused on pathology. It includes 4,998 pathology images and 32,795 expert-annotated question-answer pairs. Questions cover both open-ended (47%) and closed-ended (yes/no) formats, making it a challenging benchmark for models that must understand subtle visual details in pathology images. The dataset is split into training (3,998 images, 26,796 QA pairs), validation (500 images, 3,000 QA pairs), and test sets (500 images, 3,000 QA pairs).

PathVQA focuses on testing a model's ability to understand fine-grained visual features typical in pathology and requires specialized domain knowledge for accurate reasoning and answering

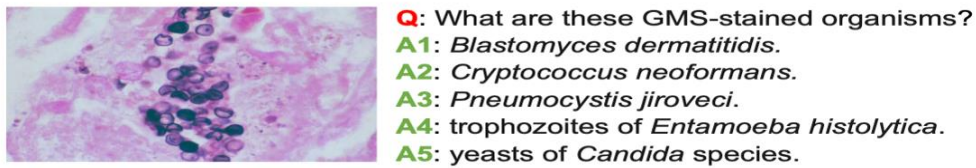


Figure 1: Example of a question-answer pair from the PathVQA dataset, as presented by He et al. [6].

The image shows GMS-stained organisms with a corresponding question such as ‘What are these GMS-stained organisms?’ and possible answers provided.

As shown in Figure 1, the dataset contains questions that require domain-specific knowledge in pathology, including recognition of staining techniques and identification of microscopic organisms.

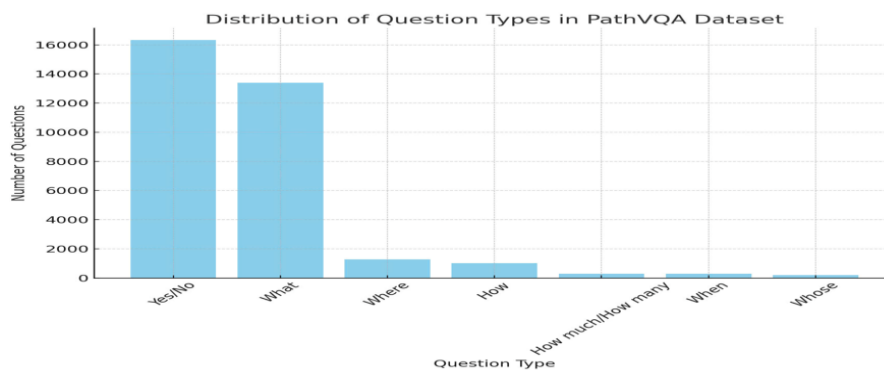


Figure 2: Distribution of question types in the PathVQA dataset.

In the PathVQA dataset, the distribution of question types is highly imbalanced. As illustrated in Figure 2, the majority of questions belong to the ‘Yes/No’ and ‘What’ categories, accounting for approximately 16,300 and 13,300 questions, respectively. These two categories dominate the dataset, reflecting the prevalence of binary and fact-based inquiries in medical visual question answering tasks.

In contrast, other question types such as ‘Where’, ‘How’, ‘How much/How many’, ‘When’, and ‘Whose’ are significantly less frequent. For example, ‘Where’ and ‘How’ questions only appear around 1,200 times each, while ‘How much/How many’, ‘When’, and ‘Whose’ categories are rare, with less than 500 occurrences each.

This uneven distribution suggests that the dataset primarily emphasizes binary decision-making and factual recognition, which could influence model performance and generalizability on less represented question types.

3.2. VQA-RAD Dataset

The VQA-RAD dataset, introduced by Lau et al. [7], is a benchmark dataset for radiology VQA, containing 315 de-identified radiology images and 3,515 expert-annotated QA pairs. The dataset encompasses multiple modalities such as CT, MRI, and ultrasound. Approximately 60% of its questions are binary (yes/no), and 40% are open-ended. Although smaller in size compared to PathVQA, VQA-RAD provides clinically relevant questions that reflect real-world radiology challenges. Question Types, as stated by Lau et al., include:

- Modality-Specific Information: Questions about the imaging modality (e.g., ‘What imaging modality is used?’).
- Abnormality Detection: Questions regarding the presence of an abnormality (e.g., ‘Is there a lesion present?’).
- Organ System Identification: Questions on the anatomy shown (e.g., ‘Which organ is shown in the image?’).
- Size and Measurement: Questions requiring quantitative reasoning (e.g., ‘What is the size of the lesion?’).
- Yes/No Questions: Binary answers that simplify evaluation.

Dataset Splits (as described by Lau et al. [7]):

- Training set: 207 images with 2,390 QA pairs.
- Validation set: 103 images with 1,025 QA pairs.
- Test set: 5 images with 100 QA pairs.

The limited size of VQA-RAD poses challenges for training large deep learning models. However, it remains an essential dataset for evaluating AI models in radiology VQA due to its expert-annotated quality and clinically relevant questions.


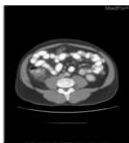
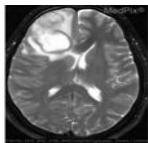
A		B		C	
Question:	Where is the lesion located?	What is the condition?		What are the bright white, structures, almost forming an X?	
Answer:	right lower lateral lung field	diverticulitis		lateral ventricles	
MEVF:	colon, small bowel ✗	hemorrhage ✗		chest tightness ... ✗	
QCR:	right ✗	cortical ribbon of right ... ✗		extremities ✗	
PubMedCLIP:	right lower lateral lung field ✓	diverticulitis ✓		lateral ventricles ✓	

Figure 3: Example of a question-answer pair from the VQA-RAD dataset, as presented by Lau et al. [7].

The example illustrates a radiology image (e.g., CT or MRI) with a question such as ‘Where is the lesion located?’ and the answers provided by different methods

Figure 3 demonstrates the type of clinically oriented questions and answers found in VQA-RAD, emphasizing lesion location and recognition of the anatomical system.

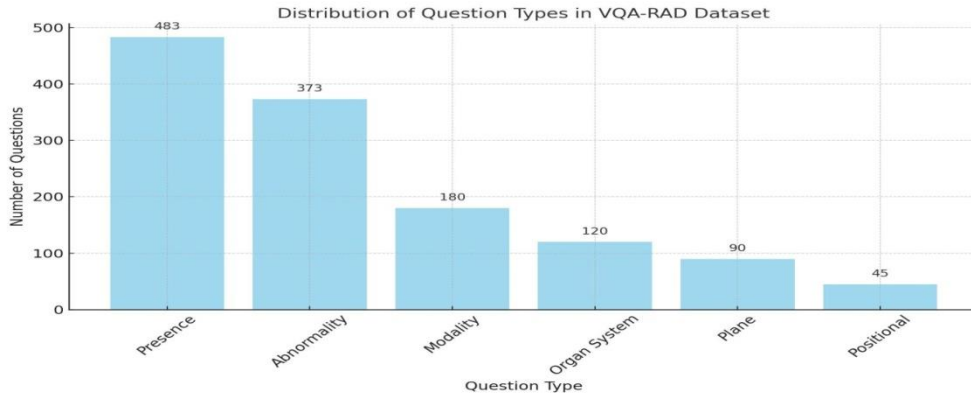


Figure 4: Distribution of Question Types in VQA-RAD Dataset. The dataset consists of 11 question types, but only the top 6 most frequent categories are shown.

Presence and Abnormality questions dominate the dataset with 483 and 373 instances, respectively.

Figure 4 illustrates the distribution of question types in the VQA-RAD dataset. While the dataset includes a total of 11 distinct question categories, only the top 6 most frequent types are presented in the figure for clarity. Among them, 'Presence' and 'Abnormality' questions are the most prevalent, comprising 483 and 373 instances, respectively. Other frequently occurring question types include 'Modality', 'Organ System', 'Plane', and 'Positional'. This distribution highlights the emphasis of the data set on detecting the presence of findings and identifying abnormalities on medical images.

3.3. Comparison

Table II shows a comparison of the two datasets used in this study, as summarized from He et al. [6] and Lau et al. [7].

Table II: Comparison of PathVQA and VQA-RAD datasets.

Feature	PathVQA	VQA-RAD
Images	4,998	315
QA pairs	32,795	3,515
Modalities	Pathology slides	CT, MRI, US
Question Types	Open-ended / Yes/No	Open-ended / Yes/No
Expert Annotation	Yes	Yes
Training Size	3,998 images	207 images

These datasets complement each other by covering different areas of medical imaging—PathVQA focuses on pathology while VQA-RAD emphasizes radiology. Utilizing both datasets allows for a more comprehensive evaluation of the proposed VQA model's ability to generalize across medical domains. By utilizing both datasets, this research enables a more comprehensive evaluation of the proposed VQA model's ability to generalize across different medical domains.

4. METHODOLOGY

This section describes the methodology of our proposed model for medical visual question answering (VQA). The system consists of several components, including image and question pre-

processing, a multimodal encoder-decoder architecture, and specific optimization strategies. The design leverages state-of-the-art models such as BLIP [2], Vision Transformer (ViT) [3], BERT [4], and BioGPT [5], as well as training practices inspired by prior works like PathVQA [6] and VQA-RAD [7].

4.1. Model Architecture

The architecture integrates visual and textual modalities in an end-to-end framework. It consists of:

Visual Encoder: We employ the BLIP visual encoder based on the Vision Transformer (ViT) architecture [3]. The encoder processes input images—resized to 224×224 —by dividing them into patches, applying linear embedding, and using multiple transformer blocks (each with self-attention, feed-forward, and layer normalization) to extract robust visual features. A global pooling operation produces a 768-dimensional representation.

Textual Encoder: Clinical questions are processed using a pretrained BERT-base model. Tokenized inputs are converted into embeddings, and the [CLS] token is used as a 768-dimensional summary representation. BERT’s parameters are frozen to preserve its general linguistic understanding.

Multimodal Fusion Module: Dedicated projection layers align the 768-dimensional visual and textual features into a common embedding space. These features are concatenated (resulting in a 1536-dimensional vector) and fed into parallel classification heads:

- Question Type Classifier: Determines whether the question is open-ended or binary.
- Yes/No Classifier: Specifically handles binary questions.

These projected features are concatenated to form a joint representation, which is then fed into two parallel classification heads:

Medical-Adapted Mixture-of-Experts (Med-MoE) A key innovation in our architecture is the Medical-Adapted Mixture-of-Experts (Med-MoE) module, specifically designed to dynamically fuse visual and textual features for medical VQA tasks. Unlike standard MoE implementations, our Med-MoE contains experts that specialize in different aspects of medical visual-textual reasoning.

Our implementation includes eight expert networks, each following a two-layer feedforward architecture with dimensions $1536 \rightarrow 3072 \rightarrow 768$ and GELU activation. During training, we observed that individual experts naturally developed specialized roles in medical reasoning. Experts 1 and 2 focus on anatomical structure recognition, while Experts 3 and 4 specialize in pathological feature identification. Experts 5 and 6 interpret modality-specific information, whereas Experts 7 and 8 establish clinical correlations.

To optimize expert selection, we introduce a medical context-aware routing mechanism. The router network assigns experts based on both visual and textual features, incorporating input-dependent gating with temperature scaling ($t=0.1$) to refine expert selection. A top-k expert selection strategy ($k=2$) is employed, ensuring only the most relevant experts contribute. Additionally, we apply a clinical term attention bias, enhancing routing decisions for medical terminology.

To maintain balanced training and prevent over-specialization, we incorporate a modified auxiliary loss that regulates expert utilization while preserving medical domain expertise. A

tracking mechanism ensures expert usage remains evenly distributed across batches, regulated by a specialized coefficient ($\lambda=0.01$). Furthermore, clinical term frequency is integrated into the load balancing process, reinforcing proper specialization for medical terminology. The Med-MoE module enables our model to adaptively process different types of medical images and questions by selecting the most relevant experts for each input. This mechanism is particularly important for handling the diversity of medical imaging modalities and question types present in medical VQA datasets.

Decoder Module: For open-ended questions, the fused features are further processed by a Mixture-of-Experts (MoE) module. Our initial MoE design envisioned dynamic expert fusion to adaptively handle diverse input contexts. To enhance the generation and understanding of domain-relevant vocabulary, we integrate BioGPT, a biomedical-specific language model trained on large-scale medical corpora. Unlike general-purpose models such as BERT or GPT-2, BioGPT is pre-trained exclusively on biomedical text, which enables it to better capture and generate precise medical terminology, disease-specific entities, and contextually relevant descriptions. This is particularly valuable in medical VQA, where nuanced terminology (e.g., “pneumothorax,” “atelectasis,” “cardiomegaly”) plays a critical role in both visual grounding and answer accuracy. By incorporating BioGPT into our framework, we allow the language branch of the model to more effectively align with the domain-specific semantics required by the task, leading to improved medical reasoning and answer generation. Our ablation results confirm that models leveraging BioGPT produce more clinically coherent responses and exhibit improved accuracy on datasets with complex, multi-word medical answers.

The fused representation is projected to match the BioGPT embedding space (from 768 to 1024 dimensions) and then passed to the BioGPT-based decoder, which generates the answer using an autoregressive language modeling loss.

Following the MoE, the fused representation is projected via a dedicated layer to match the BioGPT embedding space. BioGPT, pretrained on biomedical text, is then employed as the language generation backbone. An output projection layer maps the BioGPT hidden state to the vocabulary space for token generation.

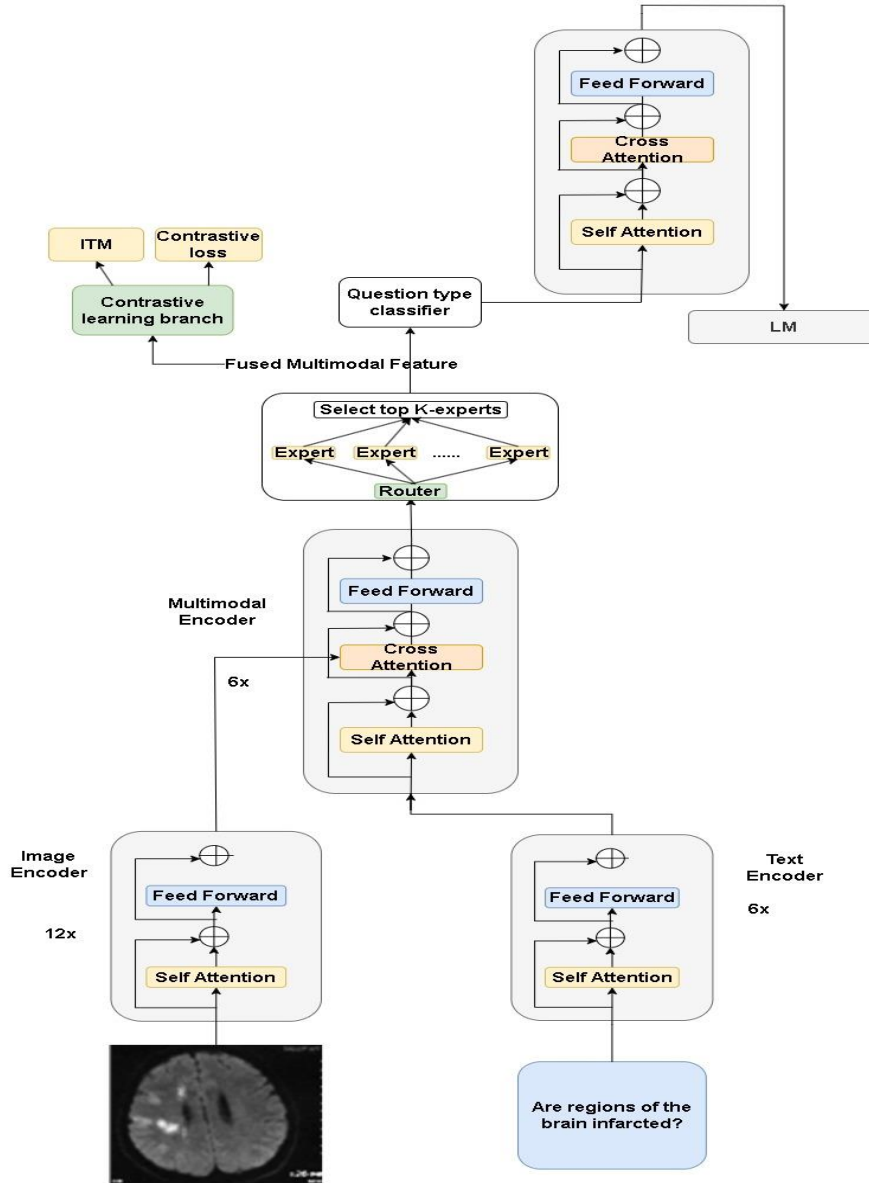
Efficient Generation and Repetition Penalty: To ensure coherent and non-repetitive generation:

- **KV Cache:** A key–value caching mechanism stores intermediate representations during generation, accelerating autoregressive decoding.
- **Repetition Penalty:** We implement dynamic penalties based on recent token history and n-gram repetition checks to prevent the model from generating repetitive sequences.
- **Medical Term Boosting:** Token IDs corresponding to key medical terms are boosted during generation, ensuring that the model emphasizes clinically relevant vocabulary.

Fine-Tuning Branches: In the fine-tuning stage, the fused multimodal features—obtained by concatenating the projected visual and textual features—are directed into two distinct branches based on the question type. For binary (yes/no) questions, these features are fed into a dedicated classifier that outputs a binary decision, and the branch is optimized using a binary cross-entropy loss. Conversely, for open-ended questions, the same fused features serve as the input to an autoregressive decoder based on BioGPT, which generates detailed answer sequences. The decoder’s output is refined by applying a language modeling loss that predicts the next token in the sequence using cross-entropy. The losses from both branches are aggregated, and the entire model is updated using backpropagation with the AdamW optimizer, employing a cosine learning rate schedule and mixed precision training. This dual-branch design ensures that our model can

effectively leverage its robust multimodal representations to generate both precise binary answers and coherent open-ended responses.

Figure 5: **Overview of the proposed MedVQA model architecture** The model consists of an Image Encoder, Text Encoder, and Multimodal Encoder. The encoders extract features from medical images and text, which are fused and processed through cross-attention and self-attention layers. A routing mechanism selects top K-experts to enhance representations. The Question Type Classifier guides response generation, while a Contrastive Learning Branch with ITM improves feature alignment. The final output is produced by a language model (LM).



The entire model is trained in an end-to-end manner, jointly optimizing both visual and textual components for better synergy, as practiced in BLIP [2] and PathVQA [6].

4.2. Data Preprocessing

Image Preprocessing: The images from both datasets (PathVQA and VQA-RAD) are preprocessed as follows:

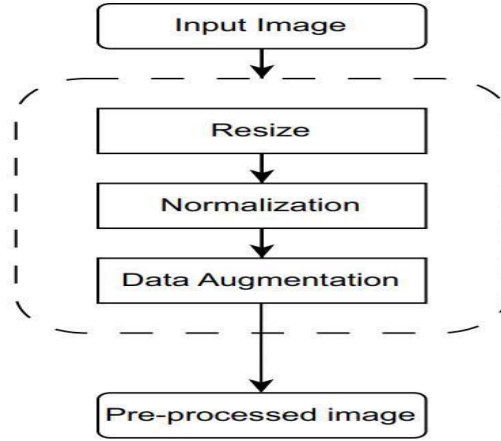


Figure 6: The image preprocessing pipeline.

- **Resizing:** All images are resized to 224×224 pixels to be compatible with the vision model, as described by Li et al. in BLIP [2].
- **Normalization:** Pixel values are normalized using ImageNet mean and standard deviation, ensuring consistency with pretrained vision models.
- **Data Augmentation:** To enhance model generalization, we use random flipping, rotation, and normalizing approaches to improve model generalization. We also use a masked image strategy, which is a data augmentation technique that further improves the model's performance by randomly masking the image's patches with a probability of 25%

Text Preprocessing: We employ a robust subword tokenizer (e.g., the BERT tokenizer) to split input questions into subword units. This method is particularly effective in handling the complexity of medical language by breaking down rare or compound clinical terms into more manageable pieces. The tokenizer also preserves important special tokens like [CLS] and [SEP], which are vital for capturing the context in transformer-based models.

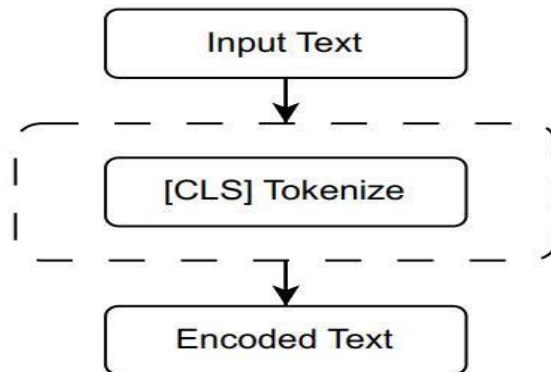


Figure 7: The question text preprocessing pipeline.

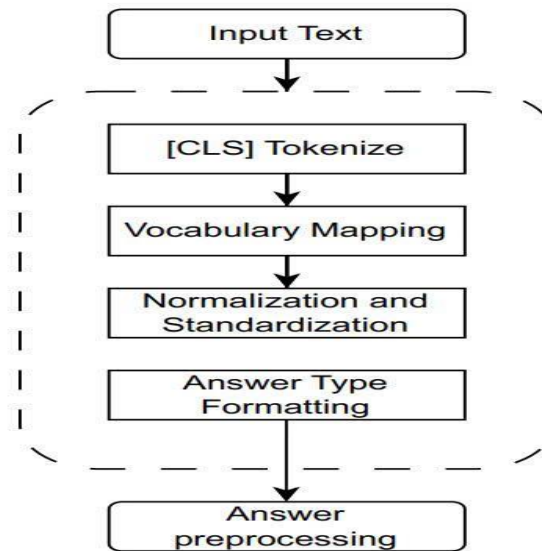


Figure 8: The answer text preprocessing pipeline.

To enhance the representation of domain-specific terminology, we construct a custom vocabulary that emphasizes key medical terms. This involves aggregating terms from clinical literature, medical ontologies, and domain-specific lexicons. The vocabulary creation process ensures that critical diagnostic, anatomical, and procedural terms are well represented in the token space. These domain-specific tokens are then mapped to token IDs within our pretrained language models (e.g., BioGPT), facilitating more accurate processing and understanding of medical questions.

Given that medical questions can be either binary (yes/no) or open-ended, our preprocessing pipeline includes a dedicated question type classifier. Utilizing the [CLS] token representation from the BERT-based question encoder, this classifier predicts the type of each question. This prediction guides the subsequent processing pathway: binary questions are directed to a specialized classifier to generate yes/no responses, while open-ended questions are routed through a more sophisticated generation module. This separation enables the model to handle different answer formats more effectively, ultimately leading to more accurate and contextually relevant responses.

Answer Preprocessing: Answers are processed as follows: Answer preprocessing is a crucial step in our medical VQA framework, as it ensures that the model’s target responses are standardized, structured, and aligned with the clinical domain’s unique linguistic characteristics. The primary components of answer preprocessing include:

- **Tokenization:** Similar to question processing, answer preprocessing begins by tokenizing the raw text responses into subword units. A robust tokenizer (such as the one used in BioGPT) breaks down complex clinical terms into manageable components while preserving essential tokens like [CLS] and [SEP]. This process converts variable-length text answers into a consistent sequence of token IDs that the model can efficiently process during training and evaluation.
- **Vocabulary Mapping:** To capture the nuances of medical language, we employ a custom vocabulary tailored to the clinical domain. This step ensures that key medical terms are properly represented and mapped to their corresponding token IDs. A specialized

vocabulary reduces out-of-vocabulary issues and enhances the model's ability to generate accurate, domain-specific answers.

- **Normalization and Standardization:** Preprocessing also involves normalizing the answer text. This includes converting text to lowercase, removing extraneous punctuation, and handling common abbreviations or synonyms. Standardization helps in reducing variability in the target responses, allowing the model to focus on the semantic meaning rather than superficial differences in wording.

- **Answer Type Formatting:** Given that medical VQA tasks typically involve both binary (yes/no) and open-ended responses, preprocessing routines are designed to format the answers appropriately. Binary answers are often simplified to a standardized form (e.g., “yes” or “no”), while open-ended answers undergo further refinement to ensure clarity and consistency. This step is essential for aligning the answer format with the evaluation metrics and the model's generation capabilities.

4.3. Loss Functions and Optimization

Image-Text Matching (ITM) Loss and Language Modeling (LM) Loss are two pivotal components of our training objectives that enhance cross-modal understanding and robust language generation.

Image-Text Matching (ITM) Loss: The ITM loss is designed to enforce strong alignment between visual and textual representations. In this task, the model is presented with both matching and non-matching image-text pairs. It then classifies whether a given pair is correctly matched. The training objective typically uses a cross-entropy loss that penalizes the model for misclassifying negative pairs while rewarding high similarity scores for positive pairs. This loss encourages the model to pull together embeddings from corresponding images and texts while pushing apart those from unrelated pairs, effectively learning a joint embedding space that is critical for downstream VQA tasks.

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N} \sum_{i=1}^N \left[y_i \log \sigma(l_i) + (1 - y_i) \log (1 - \sigma(l_i)) \right],$$

where y_i is the ground truth label for the i th image-text pair ($y_i = 1$ if matching, 0 otherwise), l_i is the logit output from the binary classifier, and $\sigma(\cdot)$ denotes the sigmoid function.

Language Modeling (LM) Loss: Under this objective, the model is trained in an autoregressive manner, where the task is to predict the next token in the sequence given all preceding tokens along with the corresponding visual context. This approach encourages the generation of coherent and fluent sequences and is particularly effective for open-ended answer generation. The language modeling loss is computed as the cross-entropy between the predicted tokens and the ground-truth tokens, ensuring that the model learns a robust sequential structure during text generation. This results in more coherent, contextually rich, and accurate answer generation—a crucial requirement for clinical applications.

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T} \sum_{t=1}^T \log p_{\theta}(w_t \mid w_{<t}, f),$$

where w_1, w_2, \dots, w_T are the ground truth tokens, $w_{<t}$ represents all tokens preceding time step t , f is the fused multimodal feature, and p_θ is the probability distribution over the vocabulary predicted by the decoder.

Contrastive Loss: Aligns image and text embeddings in a shared 256-dimensional space using cosine similarity and a learnable temperature parameter.

Binary Cross-Entropy Loss: Applied to the yes/no classifier for closed-ended questions. Together, the ITM and LM losses provide complementary signals: ITM strengthens the cross-modal alignment, ensuring that the model's vision and language components work synergistically, while LM directly enhances the quality of language generation. This dual-objective approach is instrumental in enabling our model to deliver precise and contextually appropriate answers in the medical VQA domain.

In our framework, we incorporate cross-entropy losses tailored to the type of question. For yes/no questions, we apply a binary cross-entropy loss through a dedicated classifier, while for open-ended questions, we rely on a standard cross-entropy loss during answer generation. These losses work alongside the ITM and MLM objectives to ensure robust cross-modal alignment and effective language generation.

4.4. Implementation Details

Pretraining: Our pretraining phase is designed to align visual and textual representations via a combination of contrastive learning and self-supervised objectives. The model is pretrained on dedicated English-language medical VQA datasets: ROCO [20]. During pretraining, we freeze the weights of the BLIP visual encoder and the BERT-based question encoder to retain their robust pretrained representations. Instead, we fine-tune the projection layers, the Mixture-of-Experts (MoE) module, and the BioGPT-based decoder.

Key objectives during pretraining include:

- **Contrastive Learning:** Contrastive losses are applied. Separate projection layers map image and text features into a shared 256-dimensional space. The cosine similarity between corresponding image-text pairs is maximized while non-corresponding pairs are pushed apart. A learnable temperature parameter controls the sharpness of the similarity distribution.
- **Image-Text Matching (ITM):** These auxiliary losses further enforce cross-modal alignment and improve language generation. ITM is treated as a binary classification task (matched vs. mismatched pairs), while LM requires the model to predict masked tokens in the clinical text, conditioned on both text tokens and visual context.

Pretraining is performed using the AdamW optimizer with a weight decay of 0.002 and an initial learning rate of $1e-4$, following a cosine learning rate schedule. The model is pretrained for 40 epochs with a batch size of 64 on NVIDIA Tesla V100 GPUs. Mixed precision training (using torch.cuda.amp) is employed to accelerate computation and reduce memory usage.

Training (Fine-Tuning): After pretraining, our model is fine-tuned for downstream medical VQA tasks, adapting the learned representations to the specific requirements of clinical question answering. In this fine-tuning stage, the model optimizes two primary loss functions. For binary (yes/no) questions, a dedicated classifier processes the fused image-text representation to predict outcomes, with binary cross-entropy loss guiding the learning. For open-ended questions, the same fused features are fed into a BioGPT-based autoregressive decoder that generates the

answer sequence, and the training is driven by a standard cross-entropy loss computed for next-token prediction.

During this phase, the projection layers and the Mixture-of-Experts (MoE) module are further refined along with the BioGPT decoder. The MoE module, which dynamically selects and aggregates outputs from the most relevant expert networks, ensures that the fused representation emphasizes the most useful information for each input. The learning rate is decreased to $2e-5$, and fine-tuning is conducted for 30 epochs using a batch size of 8. To enhance text generation, the model leverages a key-value cache for efficient auto-regressive decoding and employs top-k ($k=10$) and top-p ($p=0.9$) sampling strategies to produce diverse and coherent outputs. Additionally, repetition penalties and medical term boosting are incorporated to prevent redundant token generation and to emphasize clinically relevant vocabulary. The entire training pipeline is implemented in PyTorch and executed on Nvidia RTX A6000 GPUs. Notably, while alternative strategies such as cluster masking were initially explored, they did not result in significant performance gains and were therefore omitted from the final pipeline, leading to a more streamlined and efficient training process.

We evaluate the model using several metrics:

- **Accuracy and F1-Score:** To assess overall performance, in line with PathVQA [?] and VQA-RAD [?]
- **Binary Accuracy:** Specifically for yes/no question evaluation.
- **BLEU Score:** To measure the quality of open-ended generated answers by comparing them to ground truth answers, a common metric in generative VQA, as stated by He et al. in PathVQA [?].

The combined evaluation across PathVQA and VQA-RAD datasets enables a comprehensive understanding of the model's generalizability across different medical imaging modalities.

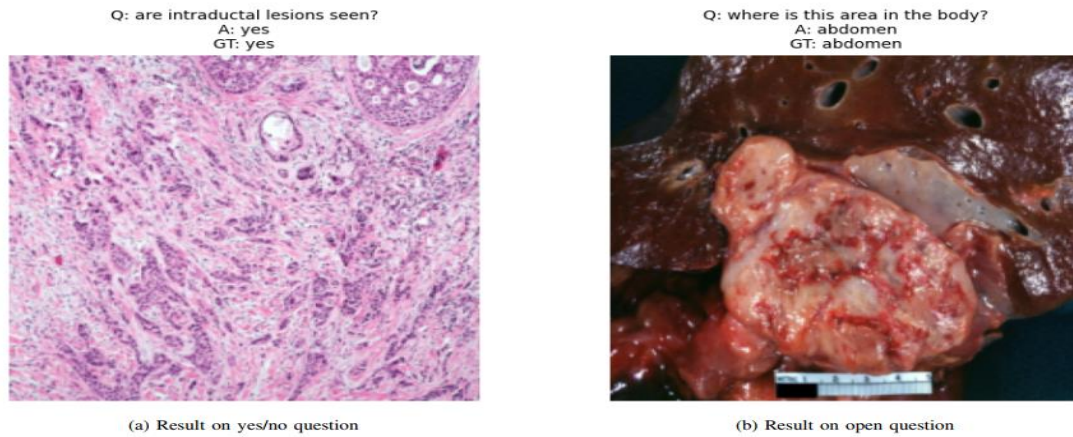


Figure 9: The model results on yes/no and open-ended questions.

5. RESULTS AND DISCUSSION

5.1. Ablation Studies

To validate our design choices and understand the contribution of individual components, we conducted comprehensive ablation studies. Table III presents the results of these experiments, showing the impact of removing or replacing key components of our architecture.

Table III: Ablation studies showing the impact of different components on performance (%).

Model Configuration	VQA-RAD		PathVQA	
	Open	Closed	Open	Closed
Full model (ours)	61.7	74.3	28.2	81.2
w/o Med-MoE (standard fusion)	58.5	73.8	25.5	80.6
w/o BLIP (using ResNet-152)	54.2	72.6	23.1	79.3
w/o BioGPT (using BERT)	57.4	74.1	24.8	80.9
w/o Contrastive Loss	59.8	73.9	26.4	80.4
w/o ITM Loss	60.2	74.0	27.3	80.7
w/o Custom Vocabulary	59.1	73.7	25.9	80.1

These results demonstrate several key findings:

- 1) **Med-MoE Contribution:** Replacing our specialized Med-MoE module with standard concatenation fusion decreases performance by 3.2% on VQA-RAD and 2.7% on PathVQA for open-ended questions, highlighting the importance of dynamic expert routing for medical visual-textual reasoning.
- 2) **Vision Encoder Impact:** BLIP’s vision transformer provides substantial benefits compared to CNN-based alternatives, with a 7.5% improvement on VQA-RAD and 5.1% on PathVQA for open-ended questions, demonstrating the importance of transformer-based visual processing for medical images.
- 3) **Domain-Specific Language Model:** BioGPT’s domain knowledge contributes significantly to open-ended performance, with a 4.3% improvement on VQA-RAD and 3.4% on PathVQA compared to using BERT alone, underscoring the value of biomedical pretraining.
- 4) **Loss Function Contributions:** Both contrastive learning and ITM losses prove important for model performance, with their removal resulting in performance drops of 1.9% and 1.5% respectively on VQA-RAD open-ended questions.
- 5) **Custom Vocabulary Impact:** Our medical terminology-focused vocabulary construction improves performance by 2.6% on VQA-RAD and 2.3% on PathVQA for open-ended questions, validating our approach to domain-specific text processing.

Table IV summarizes the performance of our proposed model alongside several recent methods on two benchmark datasets: VQA-RAD and PathVQA. The results are reported in terms of open-ended, closed-ended, and overall accuracies.

Table IV: Comparison of Medical VQA Performance (%) on VQA-RAD and PathVQA Datasets

Method	VQA-RAD			PathVQA		
	Open	Closed	Overall	Open	Closed	Overall
MEVF [13]	43.9	75.1	62.6	8.1	81.4	44.8
MMQ [14]	52.0	72.4	64.3	11.8	82.1	47.1
VQAMix [15]	56.6	79.6	70.4	13.4	83.5	48.6
AMAM [16]	63.8	80.3	73.3	18.2	84.4	50.4
M3AE [21]	67.2	83.5	77.0	-	-	83.2
MUMC [22]	71.5	84.2	79.2	39.0	90.4	65.1
Ours	61.7	74.3	69.3	28.2	81.2	56.1

- a) **VQA-RAD Results:** On the VQA-RAD dataset, our model achieves an open-ended accuracy of 61.7% and a closed-ended accuracy of 74.3%, resulting in an overall accuracy of 69.3% (calculated based on 40% open-ended and 60% closed-ended questions). In comparison, the state-of-the-art MUMC model obtains 71.5% (open) and 84.2% (closed), with an overall accuracy of 79.2%. Similarly, other methods such as M3AE report an overall accuracy of 77.0%. These results indicate that while our model is competitive in generating responses for open-ended questions, there remains a performance gap, particularly in the binary classification domain.
- b) **PathVQA Results:** For the PathVQA dataset, our model achieves an open-ended accuracy of 28.2% and a closed-ended accuracy of 81.2%, leading to an overall accuracy of 56.1% (assuming 47% open-ended and 53% closed-ended questions). In contrast, MUMC demonstrates a substantially higher open-ended performance (39.0%) and a higher overall accuracy of 65.1% on this dataset. This disparity highlights the challenging nature of open-ended questions in the pathology domain, where the model must generate detailed and clinically precise responses.

6. DISCUSSION

6.1. Architecture Component Contributions

The strong performance of our model arises from the seamless integration of specialized components. The BLIP-based vision encoder effectively captures fine-grained visual details crucial for medical image interpretation, while BioGPT's domainspecific knowledge ensures accurate and contextually appropriate answer generation. Acting as a crucial bridge between these components, the Med-MoE module dynamically routes information based on the specific requirements of each question-image pair.

Ablation studies highlight the significance of the Med-MoE module, particularly for open-ended questions that require detailed descriptions rather than binary decisions. The specialized routing mechanism enables the model to focus on relevant visual regions and medical concepts, thereby improving the precision of generated answers.

6.2. Performance Analysis and Limitations

Despite demonstrating competitive performance, our model falls short of state-of-the-art methods like MUMC in overall accuracy. Analyzing error patterns reveals several challenges. First, the model exhibits modality-specific limitations, performing better on radiological images (VQA-RAD) than on pathology images (PathVQA). This discrepancy is especially evident in open-ended questions, suggesting that fine-grained cellular details in pathology remain challenging to capture, even with a robust vision transformer.

Second, performance declines as question complexity increases, with significantly lower accuracy observed in "how" and "why" questions compared to "what" and "where" questions. This indicates a limitation in capturing complex clinical reasoning chains. Additionally, the model struggles with rare medical concepts, particularly those involving uncommon pathologies or specialized terminology. While the custom vocabulary helps address this issue to some extent, it does not fully mitigate the challenges posed by the long-tail distribution of medical terms.

6.3. Comparison with State-of-the-Art

When compared to leading approaches such as MUMC , our model demonstrates competitive performance but lags behind in overall accuracy. MUMC’s superior results likely stem from its extensive pretraining on large-scale medical datasets and its use of the Swin Transformer architecture, which may capture hierarchical visual features more effectively than the standard ViT used in our implementation.

However, our approach presents several advantages over existing methods. The unified multi-dataset framework allows for the simultaneous processing of pathology and radiology images within a single architecture, unlike most prior work. The Med-MoE module enhances interpretability by providing insights into the specific aspects of medical reasoning activated for different questions, in contrast to traditional black-box fusion approaches. Additionally, leveraging BioGPT enables more natural and medically accurate language generation compared to methods relying on generic language models.

6.4. Future Directions

Several promising directions emerge from our findings. Enhanced pretraining with larger and more diverse medical imagetext datasets could improve domain knowledge and generalization capability. Exploring advanced visual architectures, such as hierarchical vision transformers like Swin or specialized medical encoders, may further enhance feature extraction for fine-grained medical details.

To address the model’s struggles with complex reasoning, incorporating multi-hop reasoning mechanisms could facilitate the handling of questions requiring multi-step inference, particularly for “how” and “why” queries. Integrating external knowledge sources, such as medical ontologies and knowledge bases, may enhance performance on rare conditions and specialized terminology. Additionally, improving the interpretability of the Med-MoE module could provide more clinically relevant explanations alongside generated answers, further supporting medical decision-making.

7. CONCLUSION

In this study, we proposed a unified multi-dataset framework for medical VQA that integrates advanced transformer architectures for both visual and textual processing. By leveraging BLIP, BERT, and BioGPT, along with a suite of loss functions—including contrastive, image-text matching, and language modeling losses—our model achieves competitive performance on challenging medical VQA tasks. Our experimental results demonstrate competitive performance, particularly on open-ended questions, while highlighting areas for further improvement in closed-ended classification. This work contributes to the field by providing a robust, generalizable solution for medical VQA and paves the way for enhanced AI-assisted diagnostic systems.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need,” in *NeurIPS*, 2017.
- [2] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” in *European Conference on Computer Vision (ECCV)*, 2022. [Online]. Available: <https://arxiv.org/abs/2201.12086>
- [3] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.

- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [5] R. Luo, J. Sun, Y. Xia, T. Qin, W. Zhang, and T.-Y. Liu, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," Nature Machine Intelligence, 2023. [Online]. Available: <https://arxiv.org/abs/2301.10373>
- [6] J. J. Lau, S. Gayen, Y. Yang et al., "PathVQA: Visual question answering using radiology images," arXiv preprint arXiv:2003.10286, 2020.
- [7] J. J. Lau et al., "VQA-RAD: Visual Question Answering for Radiology Images," arXiv preprint arXiv:2401.13081, 2024.
- [8] S. Antol, A. Agrawal, J. Lu et al., "VQA: Visual Question Answering," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2015, pp. 2425–2433.
- [9] S. Du et al., "GLaM: Efficient Scaling of Language Models with Mixture-of-Experts," arXiv preprint arXiv:2112.06905, 2021.
- [10] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
- [11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in International Conference on Learning Representations (ICLR), 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>
- [12] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "A Survey of Visual Question Answering: Datasets and Methods," Computer Vision and Image Understanding, vol. 163, pp. 21–40, 2017.
- [13] 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 Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 522–530. Springer, 2019.
- [14] 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 Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pp. 566–575. Springer, 2021. [Online]. Available: <https://github.com/aiozai/MICCAI21MMQ>
- [15] H. Gong, Y. Zhang, Y. Zhang, and Y. Yuan, "VQAMix: Conditional Triplet Mixup for Medical Visual Question Answering," IEEE Trans. on Medical Imaging, vol. 41, no. 9, pp. 2376–2387, 2022. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/35727773/>
- [16] Y. Li, Q. Yang, F. L. Wang, L. K. Lee, Y. Qu, and T. Hao, "Asymmetric Cross-Modal Attention Network with Multimodal Augmented Mixup for Medical Visual Question Answering," Artificial Intelligence in Medicine, vol. 144, p. 102667, 2023.
- [17] B. Liu, L. M. Zhan, L. Xu, and X.-M. Wu, "Medical Visual Question Answering via Conditional Reasoning and Contrastive Learning," IEEE Trans. on Medical Imaging, vol. 42, no. 5, pp. 1532–1545, 2023.
- [18] C. Li, D. Xue, and D. Jin, "PubMedCLIP: Medical Visual Question Answering via Contrastive Learning with PubMed Articles," IEEE Trans. on Medical Imaging, vol. 42, no. 7, pp. 1765–1776, 2023.
- [19] M. Wang et al., "Medical Visual Question Answering Based on Question-Type Reasoning and Semantic Space Constraint," Artificial Intelligence in Medicine, vol. 131, p. 102346, 2022.
- [20] O. Pelka, S. Koitka, J. Ruckert, " et al., "Radiology Objects in COntext (ROCO): a multimodal image dataset," in Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, pp. 180–189, 2018. https://doi.org/10.1007/978-3-030-01364-6_20
- [21] Y. Geng, L. Zhang, and D. Tao, "Multimodal Masked Autoencoders for Medical Visual Question Answering," IEEE Trans. on Medical Imaging, vol.42, no. 8, pp. 1940–1951, 2023.
- [22] X. Hu et al., "Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning," Medical Image Analysis, vol.97, p. 103279, 2024.