Multi-modal Concept Alignment Pre-training for Generative Medical Visual Question Answering

Quan Yan, Junwen Duan[*], Jianxin Wang
Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering
Central South University
Changsha, Hunan, China
{heyq8747 and jwduan}@csu.edu.cn
jxwang@mail.csu.edu.cn

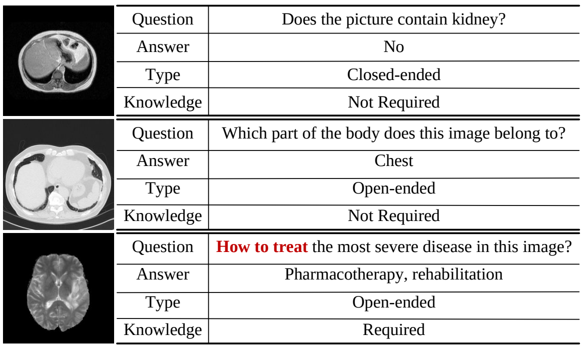
Abstract

Medical Visual Question Answering (Med-VQA) seeks to accurately respond to queries regarding medical images, a task particularly challenging for open-ended questions. This study unveils the Multi-modal Concept Alignment Pre-training (MMCAP) approach for generative Med-VQA, leveraging a knowledge graph sourced from medical image-caption datasets and the Unified Medical Language System. MMCAP advances the fusion of visual and textual medical knowledge via a graph attention network and a transformer decoder. Additionally, it incorporates a Type Conditional Prompt in the fine-tuning phase, markedly boosting the accuracy and relevance of answers to open-ended questions. Our tests on benchmark datasets illustrate MMCAP's superiority over existing methods, demonstrating its high efficiency in data-limited settings and effective knowledge-image alignment capability.

1 Introduction

Medical Visual Question Answering (Med-VQA) aims to provide accurate answers to questions about medical images, playing a crucial role in enhancing intelligent clinical services (Lin et al., 2023). Despite the growth in datasets spanning various medical conditions and anatomical areas (Lau et al., 2018; Liu et al., 2021b), Med-VQA encounters significant challenges with open-ended questions due to their complexity and the deep medical knowledge required.

Traditional approaches (Nguyen et al., 2019; Zhan et al., 2020; Gong et al., 2022) to Med-VQA have largely treated it as a classification task, drawing on predefined set of answers from training datasets, and focus on addressing data scarcity and improving cross-modal reasoning. However, these approaches struggle with the complexity of open-ended questions, which necessitate comprehension of medical knowledge, covering areas such as the treatment of a particular disease (see Figure 1) and overlook the importance of integrating external medical knowledge, a gap evident when models fail to address deeper, more nuanced questions.



Question	Does the picture contain kidney?
Answer	No
Type	Closed-ended
Knowledge	Not Required
Question	Which part of the body does this image belong to?
Answer	Chest
Type	Open-ended
Knowledge	Not Required
Question	How to treat the most severe disease in this image?
Answer	Pharmacotherapy, rehabilitation
Type	Open-ended
Knowledge	Required

Figure 1: Illustrations of Med-VQA examples, showcasing the necessity for external medical knowledge. The third example queries about the treatment of a disease, which is difficult to address without knowledge.

Recent developments in language models have shown promise in processing and generating complex text, highlighting a new direction for tackling Med-VQA challenges (Radford et al., 2019; Papanikolaou and Pierleoni, 2020; Luo et al., 2022). The key challenge now is to effectively combine these advanced language models with medical images, particularly for questions that require medical knowledge. Structuring medical concepts and their interrelations through knowledge graphs emerges as a potent strategy to enhance question answering capabilities (Fensel et al., 2020).

To address these challenges, this study regards Med-VQA as a generative task, introducing a novel approach known as **M**ulti-**m**odal **C**oncept **A**lignment **P**re-training (MMCAP). By utilizing a knowledge graph, derived from extensive medical image-caption datasets (Pelka et al., 2018; Subramanian et al., 2020) and the Unified Medical Language System(Bodenreider, 2004), MMCAP

[*]Corresponding author: Junwen Duan

Findings of the Association for Computational Linguistics: ACL 2024, pages 5378–5389
August 11-16, 2024 ©2024 Association for Computational Linguistics

achieves the alignment of images with the knowledge embedded in language model and endows the model with comprehensive image understanding capabilities. Specifically, this approach is further refined using a graph attention network (Wang et al., 2021) and a transformer decoder (Vaswani et al., 2017), facilitating the encoding of medical knowledge and its interaction with visual features. A crucial aspect of MMCAP's fine-tuning phase is the Type Conditional Prompt, which adjusts model responsiveness to the question's nature, enhancing accuracy and contextual relevance in responses.

Experimental results on two benchmark Med-VQA datasets (Liu et al., 2021b; Lau et al., 2018) demonstrate the superior performance of MMCAP. It surpasses advanced methods and achieve significant improvements in addressing open-ended questions. Further analysis demonstrates its high efficiency in utilizing limited data resources and effective knowledge-image alignment capability.

2 Related Work

2.1 Traditional Med-VQA Approaches

Initial research in Medical Visual Question Answering (Med-VQA) primarily focused on three areas: enhancing the extraction of visual features from medical images, facilitating effective cross-modal interactions between textual questions and visual data, and addressing the challenges arising from limited data availability. For instance, the MEVF framework (Nguyen et al., 2019) combines unsupervised denoising autoencoders with meta-learning strategies to significantly improve the learning of visual features. BAN-CR (Zhan et al., 2020) advances the field by incorporating bilinear attention mechanisms (Kim et al., 2018), which enable the model to differentiate and better address the demands of closed-ended and open-ended questions, resulting in improved performance on open-ended questions. Additionally, transformer-based architectures (Liu et al., 2022b, 2023c) have become increasingly popular for their ability to elevate Med-VQA performance. To tackle the issue of data scarcity, methods such as knowledge distillation (Wang et al., 2022) and data augmentation techniques (Gong et al., 2022; Li et al., 2023d) have been developed, showing promise in enhancing model robustness and performance.

2.2 Pre-training for Med-VQA

With the rapid development of pre-trained models, the Med-VQA domain has witnessed advancements through the adoption of visual-language pre-training. Image-text contrastive learning techniques (Eslami et al., 2023; Liu et al., 2021a, 2022a) have emerged as effective methods for improving feature extraction capabilities, enabling models to better understand the intricate relationship between visual and textual data in medical contexts. Furthermore, the introduction of diverse pre-training objectives (Li et al., 2023b,c; Shu et al., 2024; Chen et al., 2024) specifically designed for medical scenarios has furthered enhancements in cross-modal comprehension, thereby bolstering performance in downstream VQA tasks. Recently, with the emergence of large-scale language models (LLMs), there has been works (Zhang et al., 2023; Van Sonsbeek et al., 2023; Li et al., 2024) that utilize LLMs to improve generative question answering and extends it to Med-VQA tasks. Despite these advancements, a notable limitation of current methods is lack of emphasis on incorporating external medical knowledge. This limitation often hinders the ability of models to accurately address complex open-ended questions.

3 Method

Our methodology consists of two main phases: pre-training and fine-tuning. Initially, we focus on aligning images and external medical knowledge with textual information through generative pre-training. This phase enhances the model's understanding of medical contexts. The fine-tuning phase then adapts the pre-trained model to the Med-VQA task, leveraging the learned capability to answer medical questions accurately.

3.1 Problem Definition

Pre-training: In the pre-training phase (detailed in Section 3.2), our goal is to align the model with the rich context of medical imagery and textual information. For an input image $\mathbf{I}$ and its associated caption $\mathbf{C} = \{c_0, c_1, \ldots, c_n\}$ comprising n tokens, alongside a pre-constructed knowledge graph $\mathcal{G}$ (containing entities $\mathcal{E}$ and relations $\mathcal{R}$), the model learns to predict the next token in the caption based on the image and preceding tokens. The optimal model parameters θ^* are determined by maximizing the probability of correctly predicting

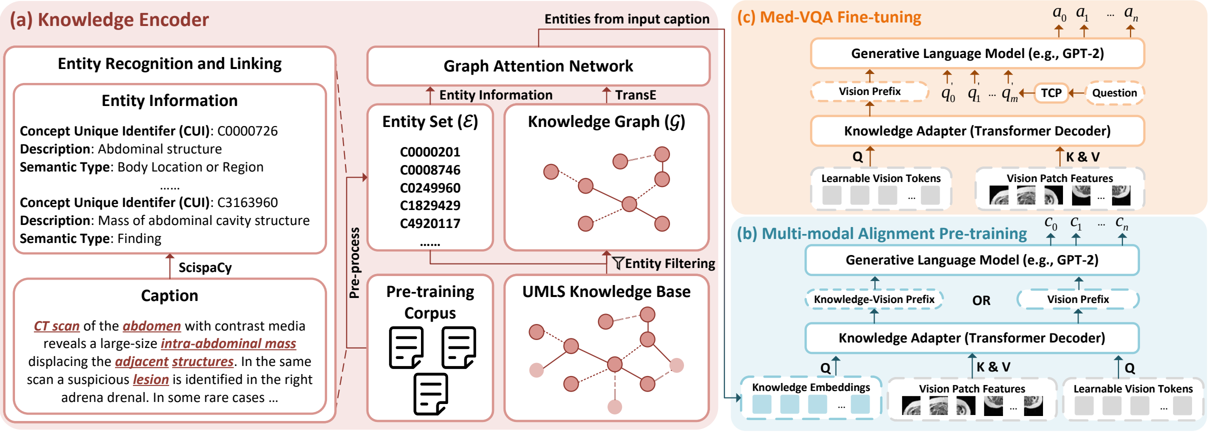

(a) Knowledge Encoder
Entities from input caption
Entity Recognition and Linking
Entity Information
Concept Unique Identifer (CUI): C0000726
Description: Abdominal structure
Semantic Type: Body Location or Region
......
Concept Unique Identifer (CUI): C3163960
Description: Mass of abdominal cavity structure
Semantic Type: Finding
ScispaCy
Caption
CT scan of the abdomen with contrast media reveals a large-size intra-abdominal mass displacing the adjacent structures. In the same scan a suspicious lesion is identified in the right adrena drenal. In some rare cases ...
Pre-process
Graph Attention Network
Entity Information
TransE
Entity Set (ℰ)
C0000201
C0008746
C0249960
C1829429
C4920117
......
Knowledge Graph (𝒢)
Entity Filtering
Pre-training Corpus
UMLS Knowledge Base
(c) Med-VQA Fine-tuning
a_0 a_1 ... a_n
Generative Language Model (e.g., GPT-2)
Vision Prefix
q_0 q_1 ... q_m
TCP
Question
Knowledge Adapter (Transformer Decoder)
Q
K & V
Learnable Vision Tokens
Vision Patch Features
(b) Multi-modal Alignment Pre-training
c_0 c_1 ... c_n
Generative Language Model (e.g., GPT-2)
Knowledge-Vision Prefix
OR
Vision Prefix
Knowledge Adapter (Transformer Decoder)
Q
K & V
Q
Knowledge Embeddings
Vision Patch Features
Learnable Vision Tokens

Figure 2: An overview of the proposed Multi-modal Concept Alignment Pre-training approach.

each token, as shown in Eq 1:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(c_i \mid \mathcal{E}, \mathcal{R}, \mathbf{I}, \mathbf{C}_{i-1}) \quad (1)$$

Fine-tuning: In the fine-tuning phase (detailed in Section 3.3), the model applies its learned alignments to answer specific medical questions. Given a medical image $\mathbf{I}$, a question $\mathbf{Q}$, and the correct answer $\mathbf{A} = \{a_0, a_1, \ldots, a_n\}$ with n tokens, the model learns to predict each token of the answer based on the image and the question. The fine-tuning process optimizes the model parameters θ^* to maximize the likelihood of the model generating the correct answer tokens, as shown in Eq 2:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(a_i \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}_{i-1}) \quad (2)$$

3.2 Multi-modal Concept Alignment Pre-training

The Multi-modal Concept Alignment Pre-training (MMCAP) approach integrates medical knowledge to enhance the understanding of medical contexts. As illustrated in Figure 2, MMCAP consists of three main components: (a) **Knowledge Encoder**, which employs a graph attention network to encode the UMLS knowledge graph, capturing external medical knowledge. (b) **Knowledge Adapter**, designed to transform vision features under the guidance of encoded knowledge, utilizing a vision encoder, learnable vision tokens, and a transformer decoder. (c) **Multi-modal Alignment Module**, leveraging a generative language model to align vision and knowledge features with texts, facilitating a coherent understanding across modalities.

3.2.1 Knowledge Graph Construction

To model external medical knowledge for multi-modal concept alignment, we construct a knowledge graph $\mathcal{G}$ from large-scale medical image-caption datasets (Pelka et al., 2018; Subramanian et al., 2020) and the unified medical language system (UMLS) (Bodenreider, 2004). As shown in Figure 2 (a), a named entity recognition, and linking tool ScispaCy[1] (Neumann et al., 2019) was applied to pre-process the captions in the pre-training corpus to link entities in the captions to the concepts (The Concept Unique Identifier, CUI) in UMLS knowledge base for entity disambiguation.

We filtered the concepts that occurred more than 10 times in the pre-training corpus. Based on the semantic types of the filtered concepts, we retained 20 semantic types most relevant to radiology, such as *"Disease or Syndrome"*, *"Body Location or Region"*, etc. Finally, a total of 15,635 medical concept entities are obtained. Based on these filtered concepts, the inter-concept relations from UMLS are introduced, such as *"has finding site"*, *"has associated finding"*, etc. These concepts and relations constitute the entity set $\mathcal{E} = \{e_1, e_2, \ldots, e_{n_e}\}$ and relation set $\mathcal{R} = \{r_1, r_2, \ldots, r_{n_r}\}$ in the knowledge graph, where n_e and n_r are the numbers of entities and relations. More details of the knowledge graph are shown in Appendix A.

3.2.2 Knowledge Encoder

The Knowledge Encoder is designed to encode entity information in the knowledge graph. Given the entity set $\mathcal{E}$, the relation set $\mathcal{R}$, and the corresponding descriptions and semantic types for each entity, the knowledge encoding process is formulated into

[1] The model of ScispaCy used in this study is *scibert-base*.

four steps: (i) A knowledge embedding learning algorithm (e.g., TransE (Bordes et al., 2013)) is applied to $\mathcal{E}$ and $\mathcal{R}$ to obtain the entity embeddings $\mathbf{f}^{ENT} \in \mathbb{R}^{n_e \times d_k}$ and relation embeddings $\mathbf{f}^{REL} \in \mathbb{R}^{n_r \times d_k}$, where d_k is the TransE embedding size. (ii) A transformer-based text encoder (e.g., BERT (Gu et al., 2021)) is applied to extract the description embedding $\mathbf{f}^{DES} \in \mathbb{R}^{n_e \times d_t}$ and semantic type embedding $\mathbf{f}^{STY} \in \mathbb{R}^{n_e \times d_t}$, where d_t is the text embedding size. Entity descriptions and semantic types are specialized interpretations of medical concepts, and such information helps to integrate medical knowledge. (iii) The nodes representation $\mathbf{f}^{NODE}$ of the knowledge graph is obtained via fusing the TransE embedding and the two types of text embeddings, as illustrated in Eq. 3,

$$\mathbf{f}^{NODE} = Norm([\mathbf{f}^{ENT} \| (\mathbf{f}^{DES} + \mathbf{f}^{STY})]) \quad (3)$$

where $Norm(\cdot)$ and $\|$ represent the layer normalization and concatenate operator. (iv) Finally, an Edge-featured Graph Attention Network (Wang et al., 2021) (EGAT) is applied to take into account the whole structure of the graph by aggregating local information for each node. EGAT incorporates edge features into node interaction, which helps obtain fine-grained graph representation via the multiple relations between nodes. The message propagation of EGAT is illustrated in Eq. 4- 5,

$$h_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} W^{(l)} h_j^{(l)} \quad (4)$$

$$\alpha_{i,j} = \text{softmax}_i(\text{LeakyReLU}(A[h_i \| e_{ij} \| h_j])) \quad (5)$$

where $h_i^{(l)}$ and $W^{(l)}$ are the features of the i-th node and the weight matrix at the l-th layer, $e_{ij} \in \mathbf{f}^{REL}$ and α_{ij} are the edge features and attention weight between the i-th and the j-th node, $\mathcal{N}(i)$ is the neighbor node set of the i-th node. After the message propagation of EGAT, the enriched knowledge embeddings $\mathbf{f}^{K} \in \mathbb{R}^{n_e \times d}$ are obtained and then used as the input of the Knowledge Adapter.

3.2.3 Knowledge Adapter

The Knowledge Adapter serves two primary functions: transforming vision features into the language model's embedding space and adapting these features with knowledge embeddings to inject medical knowledge. Following the approach of previous studies (Liu et al., 2023a; Li et al., 2023a), we utilize a 12-layer transformer decoder as the mechanism for adjusting vision features.

Initially, for each image-caption pair in the pre-training datasets, a vision encoder like CLIP (Radford et al., 2021) extracts vision patch features $\mathbf{f}^{I} \in \mathbb{R}^{l_x \times d}$ from the image, where l_x represents the length of the vision features. Concurrently, ScispaCy identifies entities within the caption, selecting corresponding embeddings $\mathbf{f}^{K}_{select} \in \mathbb{R}^{l_k \times d}$ from the knowledge embeddings, with l_k indicating the number of entities identified.

As shown in Figure 2 (b), the Knowledge Adapter incorporates two pathways to achieve its goals. For the vision feature transformation, several learnable tokens $\mathbf{f}^{V} \in \mathbb{R}^{l_v \times d}$ are set as the query input to adjust vision features, where l_v is the length of learnable tokens. Through the adjustment of learnable tokens, vision features are transformed into vision prefix $\tilde{\mathbf{f}}^{V} \in \mathbb{R}^{l_v \times d}$. For the knowledge injection, the selected knowledge embeddings $\mathbf{f}^{K}_{select}$ are as the query input to adjust vision features. In this way, vision features are adjusted into knowledge-vision prefix $\tilde{\mathbf{f}}^{K} \in \mathbb{R}^{l_k \times d}$ under the guidance of knowledge. Subsequently, vision and knowledge-vision prefix can be used as inputs of the language model, respectively, and aligned with the corresponding text.

3.2.4 Multi-modal Alignment

In order to align the images and knowledge with the language model, the two output prefixes from the Knowledge Adapter will be linearly transformed as the inputs to the language model, which is an auto-regressive generation model, in this case GPT-2-PubMed (Radford et al., 2019). Specifically, $\tilde{\mathbf{f}}^{V}$ and $\tilde{\mathbf{f}}^{K}$ are used as prefix inputs to independently generate corresponding caption $\mathbf{C} = \{c_0, c_1, \ldots, c_n\}$. At each time step i, the output is the logits parametrizing categorical distribution $p^{v}_{\theta}(\mathbf{C})$ and $p^{k}_{\theta}(\mathbf{C})$ over the vocabulary tokens, as illustrated in eq. 6- 7.

$$\log p^{v}_{\theta}(\mathbf{C}) = \sum_{n} \log p^{v}_{\theta}(c_i \mid \tilde{\mathbf{f}}^{V}, \mathbf{C}_{i-1}) \quad (6)$$

$$\log p^{k}_{\theta}(\mathbf{C}) = \sum_{n} \log p^{k}_{\theta}(c_i \mid \tilde{\mathbf{f}}^{K}, \mathbf{C}_{i-1}) \quad (7)$$

After pre-training, the language model possesses the ability to generate corresponding text based on images and knowledge and possesses a deep comprehension of images and knowledge.

3.3 Fine-tuning on Med-VQA Datasets

Upon completion of the pre-training, we retain the vision encoder for vision feature extraction, the knowledge adapter for vision feature transformation, and the language model for generative VQA. These components, collectively developed during pre-training are integrated into the subsequent fine-tuning for the downstream Med-VQA task.

3.3.1 Type Conditional Prompt

In related work (Zhan et al., 2020; Liu et al., 2022a), it has been demonstrated that incorporating reasoning conditioned on the question type can enhance Med-VQA. In this paper, we introduce a novel approach known as Type Conditional Prompt (TCP), designed to emphasize the question type through a specific prompt. Specifically, when presented with an input question $\mathbf{Q}$, a classifier is employed to discern the answer type (Closed-ended or Open-ended) and content type (Organ, Modality, Abnormality, etc.) associated with $\mathbf{Q}$. Subsequently, the identified type information is integrated into a pre-defined template, as exemplified by: *"**Open-ended** question about **Organ**: What is the function of the organ on the top of this image? The answer is:"*. The resulting filled prompts are denoted as $\mathbf{Q}^{'} = \{q_0^{'}, q_1^{'}, \ldots, q_m^{'}\}$ and are utilized as the input of the language model. More details of type conditional prompt are available in Appendix B.

3.3.2 Generative Med-VQA

For the generative Med-VQA, as illustrated in Figure 2 (c), given an input image $\mathbf{I}$ and a question $\mathbf{Q}$, the visual encoder and knowledge adapter are employed to obtain the vision prefix $\tilde{\mathbf{f}}^V$, and the question classifier is used to build the prompt $\mathbf{Q}^{'}$. Subsequently, the vision prefix and the embedded prompt are concatenated as the input of the language model to generate the answer $\mathbf{A} = \{a_0, a_1, \ldots, a_n\}$ token by token. This process is expressed in Eq. 8.

$$\log p_\theta^a(\mathbf{A}) = \sum_n \log p_\theta^a(a_i \mid \tilde{\mathbf{f}}^V, \mathbf{Q}^{'}, \mathbf{A}_{i-1}) \quad (8)$$

4 Experiment

4.1 Datasets and Metrics

We perform the Multi-modal Concept Alignment Pre-training across the following two datasets.

- **ROCO** (Pelka et al., 2018) contains over 81,000 radiology images with multiple medical imaging modalities. All images in ROCO have corresponding captions.
- **MedICaT** (Subramanian et al., 2020) is a medical image dataset in context, which consists of 217,000 images with captions from 131,000 open-access biomedical papers.

For fine-tuning and evaluation, we utilize:

- **VQA-RAD** (Lau et al., 2018) contains 315 radiology images and 3,515 question-answer pairs, with 3,064 pairs for training and 451 pairs for testing. There may be multiple question types of questions regarding a radiology image such as "modality", "abnormality", etc.
- **SLAKE** (Liu et al., 2021b) is a bi-lingual Med-VQA dataset. In this paper, we use the English version of SLAKE, which contains 642 radiology images, 7,033 question-answer pairs. Following the original splitting, where 4,919 pairs are used for training, 1,053 pairs for validation, and 1,061 pairs for testing.

Evaluation metrics align with previous work, which reflect the model's ability to handle both the breadth of medical knowledge (open-ended accuracy) and specificity (closed-ended accuracy), along with an overall Q&A accuracy metric that provides a holistic view of performance.

4.2 Implementation Details

The architecture of our Multi-modal Concept Alignment Pre-training (MMCAP) model integrates a ResNet-based CLIP encoder (Radford et al., 2021) for visual encoding and BioMedBERT (Gu et al., 2021) for textual encoding. The incorporation of GPT-2-PubMed as the language model (Papanikolaou and Pierleoni, 2020), pre-trained on extensive medical literature corpus, is intended to augment the model's comprehension of medical context.

For the Knowledge Encoder, we employ TransE embeddings of size 256, coupled with a dual-layer EGAT to enhance the relational knowledge integration. The Knowledge Adapter is designed with 16 vision tokens and a stacked transformer decoder with 1,024 feature size dimensions and 12 layers.

During pre-training, we filter non-radiology samples in the ROCO and MedICaT datasets, retaining about 200,000 image-caption pairs. The whole model is optimized using a batch size of 8, which is accumulated over two batches for efficiency, and employ an initial learning rate of 1e-4, which undergoes a cosine decay after 5,000 batches warm-up, spanning across 125,000 optimization steps.

During fine-tuning, we freeze the parameters of visual feature extractor to retain the image understanding capabilities acquired during pre-training. The batch size is adjusted to 16. The learning rate maintains the value of 1e-4, decaying according to a cosine curve over 30,000 steps for the SLAKE dataset and 18,000 steps for the VQA-RAD dataset.

Table 1: Comparison of the proposed MMCAP with state-of-the-art methods. The Q&A accuracy of closed-ended questions, open-ended questions, and all questions are reported respectively.

Method	**SLAKE**			**VQA-RAD**		
	Open	Closed	Overall	Open	Closed	Overall
MEVF+BAN+CR (Zhan et al., 2020)	-	-	-	60.0	79.3	71.6
CPRD+BAN (Liu et al., 2021a)	81.2	83.4	82.1	61.1	80.4	72.7
CP+BAN+CR (Liu et al., 2022a)	80.5	84.1	81.9	60.5	80.4	72.5
MQAT (Liu et al., 2022b)	79.7	87.7	82.8	49.8	76.3	65.7
VQAMix (Gong et al., 2022)	-	-	-	56.6	79.6	70.4
MHKD-MVQA (Wang et al., 2022)	-	-	-	63.1	80.5	73.6
MPR (Ossowski and Hu, 2023)	78.3	84.9	80.9	60.5	81.6	73.2
PubMedCLIP (Eslami et al., 2023)	78.4	82.5	80.1	60.1	80.0	72.1
ACMA-MAM (Li et al., 2023d)	80.8	86.7	83.1	63.6	**84.4**	76.1
VQA-Adapter (Liu et al., 2023b)	79.2	83.7	81.0	66.1	82.3	75.8
MITER (Shu et al., 2024)	79.2	84.4	81.2	59.4	80.5	72.1
M^{3}AE (Chen et al., 2024)	80.3	87.8	83.2	67.2	83.5	77.0
MMCAP (Ours)	**82.8**	**88.0**	**84.8**	**70.0**	83.1	**77.8**

This approach is meticulously designed to tailor the pre-trained model to the Med-VQA tasks.

The complete two-stage experiment costs about 36 hours on an NVIDIA GeForce RTX3090 GPU.

4.3 Comparison with Advanced Methods

Before comparing MMCAP with the most advanced methods, we will succinctly introduce these methods for comparison:

Methods without pre-training:

- **MEVF+BAN+CR** (Zhan et al., 2020) is a framework containing a question-conditioned and a type-conditioned reasoning module.
- **CPRD+BAN** (Liu et al., 2021a) is a framework by transfer learning and distilling a lightweight visual feature extractor.
- **CP+BAN+CR** (Liu et al., 2022a) applies a visual feature extractor via contrastive learning and a conditional reasoning framework.
- **MQAT** (Liu et al., 2022b) is an improved transformer-based model for Med-VQA.
- **VQAMix** (Gong et al., 2022) is a data augmentation method, which generates training samples by linearly combining VQA samples.
- **MHKD-MVQA** (Wang et al., 2022) applies multi-modal hierarchical knowledge distillation for Med-VQA.
- **MPR** (Ossowski and Hu, 2023) is a generative model that integrates retrieved prompts and multi-modal features to generate answers.

Methods with visual-language pre-training:

- **PubMedCLIP** (Eslami et al., 2023) is a fine-tuned version of CLIP for the medical domain based on PubMed articles.
- **ACMA-MAM** (Li et al., 2023d) constructs an image-guided attention and a question-guided attention to improve multi-modal interactions.
- **VQA-Adapter** (Liu et al., 2023b) is a parameter efficient adapter component, in which only the light-weight adapter needs to be tuned.
- **MITER** (Shu et al., 2024) is a joint adaptive pre-training framework for Med-VQA via multi-level contrastive learning.
- **M^{3}AE** (Chen et al., 2024) is a self-supervised learning paradigm, which learns to map medical images and texts to a joint space by reconstructing pixels and tokens.

The experimental results of the proposed MMCAP on SLAKE and VQA-RAD datasets are presented in Table 1. Compared with strong competitors, MMCAP achieves better performance on both datasets. Notably, MMCAP exhibits substantial improvements in the accuracy of open-ended questions, outperforming the best competitor M^{3}AE (Chen et al., 2024) by 3.1% and 4.2%, respectively. These results serve as compelling evidence for the efficacy of the introduced method.

4.4 Further Analysis

4.4.1 Performance with Limited Data

This experiment aims to evaluate MMCAP's adaptability in scenarios with varying amounts of do-

Table 2: Ablation experimental results for each proposed module. The BLEU-1 metric measuring the similarity of generated and reference answers is additionally reported.

ID	Setting				SLAKE				VQA-RAD			
	Img-Cap	K-Adp	EGAT	TCP	BLEU-1	Open	Closed	Overall	BLEU-1	Open	Closed	Overall
1	✗	✗	✗	✗	77.4	77.8	83.1	79.9	50.4	49.7	75.0	65.0
2	✗	✗	✗	✔	79.2	80.2	82.9	81.2	54.9	55.9	75.7	67.9
3	✔	✗	✗	✔	79.1	79.7	87.3	82.7	70.9	67.6	77.2	73.4
4	✔	✔	✗	✔	80.1	80.8	87.7	83.5	73.6	66.5	82.0	75.8
5	✔	✔	✔	✔	**81.7**	**82.8**	**88.0**	**84.8**	**75.6**	**70.0**	**83.1**	**77.8**

main data. By systematically reducing the data volume and observing the impact on model performance, we can assess the robustness of MMCAP. The settings include: without pre-training, without knowledge, and MMCAP. Figure 3 presents the experimental results on the SLAKE dataset.

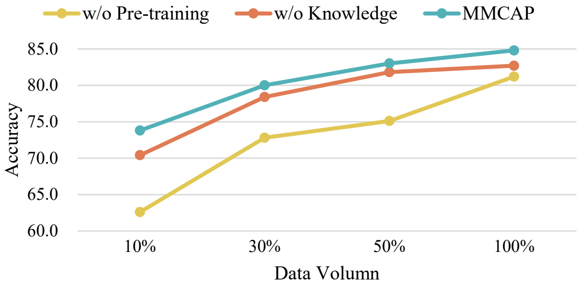


Figure 3: Results of fine-tuning experiments on the SLAKE dataset with limited data. Overall accuracy is reported and the horizontal coordinates represent the percentage of data volume used.

Without pre-training, the model exhibits a noticeable decrease in performance with the reduced data volume. However, following image-text generation pre-training (without knowledge), the model demonstrates improved learning capabilities for limited data, owing to enhanced image comprehension. Introducing the proposed Knowledge Encoder and Adapter further enhances the model's performance with limited data. Notably, the accuracy achieved is comparable to the best competitor M^{3}AE (Chen et al., 2024) when utilizing only 50% of the training data (83.0 vs. 83.2).

These results validate the model's capability to leverage limited training data, showcasing its potential in scenarios with reduced data availability.

4.4.2 Performance with Different LMs

Exploring different language models allows us to pinpoint the importance of domain-specific pre-training. In this section, we evaluate the performance of various fine-tuned language models on the Med-VQA task, including four language models: BioGPT (Luo et al., 2022), GPT-2 (Radford et al., 2019), GPT-2-Medium, and GPT-2-PubMed-Medium (Gu et al., 2021). As shown in Figure 4, the results reveal that GPT-2 outperforms BioGPT on the Med-VQA task. Notably, GPT-2-PubMed, pre-trained on medical corpus, exhibits the highest performance. This observation suggests that MMCAP's ability to enhance the Med-VQA task is particularly pronounced when the underlying language model possesses richer medical domain knowledge. The incorporation of medical domain-specific pre-training, as demonstrated by GPT-2-PubMed, contributes significantly to the overall improvement in Med-VQA performance.

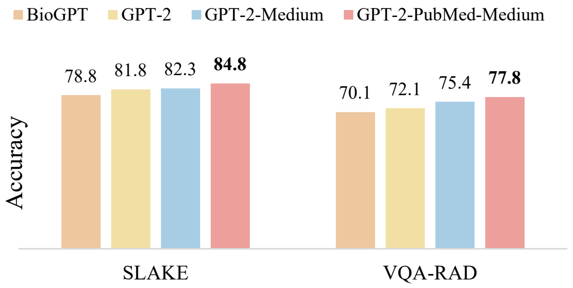


Figure 4: Results of the experiments on different language models, with overall accuracy reported.

4.4.3 Ablation Study

To assess the individual contributions of the modules proposed in this study, we conducted ablation studies on four key components: image-caption generation (Img-Cap), knowledge adapter (K-Adp), graph attention network (EGAT), and type conditional prompt (TCP). The results of the ablation experiments are presented in Table 2.

The comparison between the first and second rows demonstrates the effectiveness of the proposed TCP, showcasing its ability to enhance Q&A

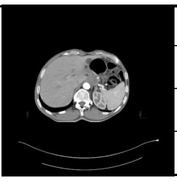
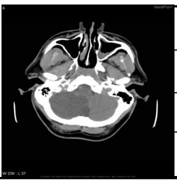
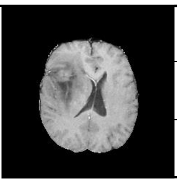
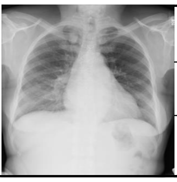
Improved Cases		**Question**	What abnormality is present?
		Answer	**Bleeding in** right posteroinferior cerebellum
		w/o Knowledge	**Right posteroinferior cerebellum**
		MMCAP	**Bleeding in** right posteroinferior cerebellum
		**Question**	What is the effect of the organ on the upper left of this image?
		Answer	**Biotransformation detoxification**
		w/o Knowledge	**Improve body's immunity**
		MMCAP	**Biotransformation detoxification**
Erroneous Cases		**Question**	What diseases are included in the picture?
		Answer	**Brain edema brain non enhancing tumor**
		MMCAP	**Brain edema brain enhancing tumor brain non enhancing tumor**
		**Question**	Where is/are the abnormality located?
		Answer	**Right lung left**
		MMCAP	**Right lung upper left**

Figure 5: Two case studies of medical visual question answering, demonstrate the improvements and limitations of MMCAP on open-ended questions, respectively.

accuracy by emphasizing question types.

Subsequently, comparing the second and third rows reveals that the pre-training of image-caption generation positively impacts model performance. This objective aligns the images with the corresponding captions, enhancing the model's image comprehension. However, without the injection of knowledge, the accuracy remains suboptimal.

In the experimental settings of the third, fourth, and fifth rows, we incrementally introduced the knowledge adapter and the graph attention network. Notably, in the absence of EGAT, the improvement of the model on open-ended questions is still limited, resulting in unsatisfactory overall accuracy. However, the addition of EGAT integrates the information of relevant medical concepts beyond captions through graph aggregation. In this way, the knowledge adapter can receive richer medical knowledge and brings more significant improvements on open-ended questions.

4.4.4 Case Study

To comprehensively assess the strengths and limitations of MMCAP in the domain of medical visual question answering, we present two sets of test cases illustrated in Figure 5.

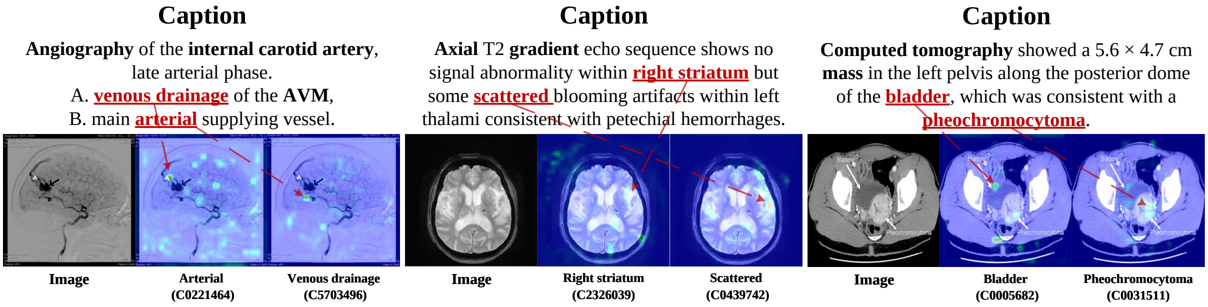
In the improved cases, the first case involves a typical open-ended question concerning abnormalities. Without medical knowledge injection, the model struggles to recognize the most critical symptoms of bleeding. The second case involves an open-ended question requiring medical knowledge, pertaining to the effects of specific organs, demanding intricate reasoning capabilities, pose a challenge without knowledge intervention. However, when MMCAP is augmented with external medical knowledge, it adeptly answers these questions correctly.

In the erroneous cases, which focus on disease presence and location, the first case reveals that while MMCAP identified the diseases mentioned in the reference answer, it additionally generated an non-existent disease “brain enhancing tumor”. In the second case, MMCAP failed to provide the most accurate location, stating “right lung left”.

The above cases prove that the proposed method helps to address open questions that require medical knowledge, but still has limitations in multi-disease detection and disease location recognition.

4.4.5 Visualization

To validate the effectiveness of the proposed Knowledge Adapter, i.e., its ability to align images with knowledge, we presents a set of visualization results related to the pre-training process. The attention heatmap between knowledge embeddings



Angiography of the **internal carotid artery**, late arterial phase. A. **venous drainage** of the **AVM**, B. main **arterial** supplying vessel.

Axial T2 **gradient** echo sequence shows no signal abnormality within **right striatum** but some **scattered** blooming artifacts within left thalami consistent with petechial hemorrhages.

Computed tomography showed a 5.6 × 4.7 cm **mass** in the left pelvis along the posterior dome of the **bladder**, which was consistent with a **pheochromocytoma**.

Figure 6: Visualization of attention heatmap between knowledge embeddings and image features.

and visual features is visualized to verify whether the knowledge embeddings focus on the image regions relevant to medical entities when adjusting visual features. As shown in Figure 6, the upper side displays caption corresponding to the image, where the parts marked in red represent medical concepts, corresponding to the attention heatmap visualization below. For instance, in Case 1, the embeddings of *"venous drainage"* and *"artery"* can focus on the respective vessel positions marked in the image; in Case 2, the embedding of *"scattered"* can focus on the location of the bleeding point on the right side of the image; and in Case 3, the embeddings of *"bladder"* and *"pheochromocytoma"* can also focus on the corresponding positions in the image. Such visualization analysis vividly demonstrates the effectiveness of the knowledge adapter in promoting the correspondence between entities in the knowledge graph and images, thereby achieving more precise concept alignment.

5 Conclusion

In this work, we redefined Med-VQA as a generative task and design a Multi-modal Concept Alignment Pre-training (MMCAP) method based on the specifics and shortcomings of current methods. With a constructed medical knowledge graph and a knowledge alignment pre-training method, MMCAP surpasses existing methods and achieves significant improvements on open-ended questions. Further analysis demonstrates its high efficiency in utilizing limited data resources and effective knowledge-image alignment capability.

Limitations

The proposed approach has several limitations: (i) MMCAP is based on knowledge-enhanced generative vision-language pre-training, and although it achieves significant improvements in medical visual question answering, however, its potential for other cross-modal tasks in medicine, such as radiology report generation, image-text retrieval, etc., has yet to be explored. (ii) Despite the improvements achieved in addressing open-ended questions that require medical knowledge, the sensitivity of the current approach to the local position of medical images is still limited, leading to subtle biases in individual questions about position.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No.2021YFF1201200), the Science and Technology Major Project of Changsha (No.kh2402004). This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2024. Mapping medical image-text to a joint space via masked modeling. *Medical Image Analysis*, 91:103018.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan

Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. 2020. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10.

Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. 2022. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11):3332–3343.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. 2023b. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023c. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Yong Li, Qihao Yang, Fu Lee Wang, Lap-Kei Lee, Yingying Qu, and Tianyong Hao. 2023d. Asymmetric cross-modal attention network with multimodal augmented mixup for medical visual question answering. *Artificial Intelligence in Medicine*, 144:102667.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021b. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. 2022a. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging*, 42(5):1532–1545.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. 2023b. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Lei Liu, Xiangdong Su, Hui Guo, and Daobin Zhu. 2022b. A transformer-based medical visual question answering model. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1712–1718. IEEE.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023c. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pages 445–456. Springer.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer.

Timothy Ossowski and Junjie Hu. 2023. Multimodal prompt retrieval for generative visual question answering. *arXiv preprint arXiv:2306.17675*.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Chang Shu, Yi Zhu, Xiaochu Tang, Jing Xiao, Youxin Chen, Xiu Li, Qian Zhang, and Zheng Lu. 2024. Miter: Medical image–text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications*, 238:121526.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*.

Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–736. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jianfeng Wang, Shuokang Huang, Huifang Du, Yu Qin, Haofen Wang, and Wenqiang Zhang. 2022. Mhkd-mvqa: Multimodal hierarchical knowledge distillation for medical visual question answering. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 567–574. IEEE.

Ziming Wang, Jun Chen, and Haopeng Chen. 2021. Egat: Edge-featured graph attention network. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30*, pages 253–264. Springer.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

A Details of Knowledge Graph

To ensure the reproducibility of MMCAP, we provide the compositional structure of the knowledge graph constructed in this section. Our knowledge graph constructed includes 15,635 medical entity nodes covering 20 semantic types, where there are 130,196 edges between entities covering a total of 50 relation types from UMLS. Table 3 shows the details of the semantic types and relation types.

B Details of Type Conditional Prompt

This section supplements the details of the proposed type conditional prompt. Figure 7 illustrates the distribution of question content types in SLAKE and VQA-RAD datasets, where the question classifiers are pre-trained based on dataset-specific content types. Table 4 demonstrates an example of prompt construction. In this case, the question classifier will recognize the answer type *"Open-ended"* and the content type *"Position"* of the input question $\mathbf{Q}$, and fills in the template $\mathbf{T}$ to obtain the type-conditional prompt $\mathbf{Q}^{'}$.

Table 3: The compositional structure of the constructed knowledge graph, covering 20 semantic types of entities, and 50 relations.

Semantic Types	1.Body Part, Organ, or Organ Component, 2.Disease or Syndrome, 3.Finding, 4.Gene or Genome, 5.Therapeutic or Preventive Procedure, 6.Neoplastic Process, 7.Pharmacologic Substance, 8.Diagnostic Procedure, 9.Body Location or Region, 10.Spatial Concept, 11.Pathologic Function, 12.Medical Device, 13.Qualitative Concept, 14.Body Space or Junction, 15.Congenital Abnormality, 16.Quantitative Concept, 17.Cell Component, 18.Injury or Poisoning, 19.Functional Concept, 20.Sign or Symptom.
Relations	1.inverse isa, 2.isa, 3.has finding site, 4.finding site of, 5.associated morphology of, 6.has associated morphology, 7.same as, 8.possibly equivalent to, 9.method of, 10.has method, 11.has manifestation, 12.manifestation of, 13.part of, 14.has part, 15.disease has associated anatomic site, 16.is associated anatomic site of, 17.laterality of, 18.has laterality, 19.has direct procedure site, 20.direct procedure site of, 21.pathological process of, 22.has pathological process, 23.use, 24.used for, 25.anatomic structure is physical part of, 26.has physical part of anatomic structure, 27.prev symbol of, 28.has prev symbol, 29.related to, 30.clinically similar, 31.regional part of, 32.has regional part, 33.is primary anatomic site of disease, 34.disease has primary anatomic site, 35.disease may have finding, 36.may be finding of disease, 37.is finding of disease, 38.disease has finding, 39.is not primary anatomic site of disease, 40.disease excludes primary anatomic site, 41.has location, 42.location of, 43.associated finding of, 44.has associated finding, 45.is location of anatomic structure, 46.anatomic structure has location, 47.disease may have associated disease, 48.disease excludes finding, 49.may be associated disease of disease, 50.is not finding of disease.

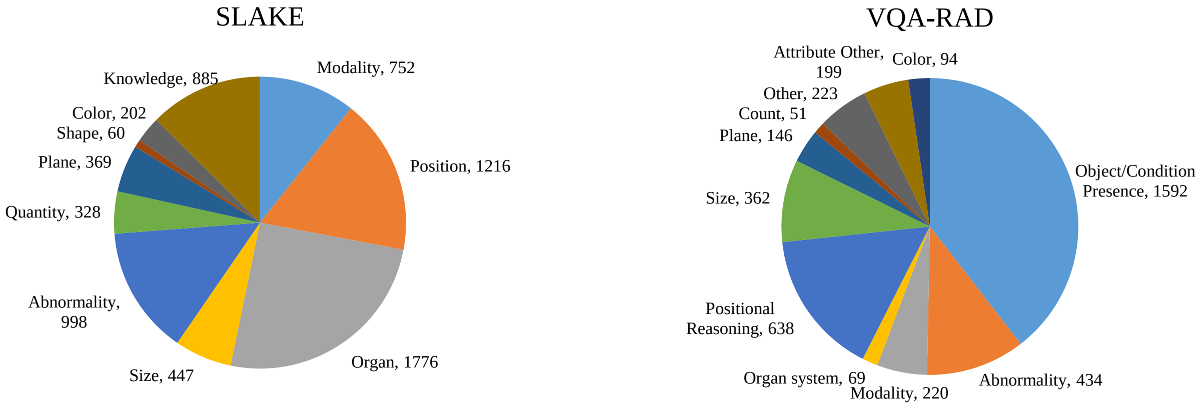


Figure 7: Distribution of question content types in SLAKE and VQA-RAD datasets.

Table 4: Example of type conditional prompt construction.

Input Question $\mathbf{Q}$	Where is the abnormality in this image?
Answer Type	Open-ended
Content Type	Position
Template $\mathbf{T}$	{**Answer Type**} question about {**Content Type**}: {**Question**}, the answer is:
Prompt $\mathbf{Q}'$	**Open-ended** question about **Position**: **Where is the abnormality in this image?** The answer is: