Seek Inner: LLM-Enhanced Information Mining for Medical Visual Question Answering

Ao Ma
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
aoma0081@uni.sydney.edu.au

Zhiyuan Li
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
zhli0736@uni.sydney.edu.au

Zhuonan Liang
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
zhuonan.liang@sydney.edu.au

Tiancheng Gu
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
tigu8498@uni.sydney.edu.au

Jianan Fan
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
jianan.fan@sydney.edu.au

Jieting Long
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
jlon5443@uni.sydney.edu.au

Henning Müller*†
The University of Applied Sciences
Western Switzerland (HES-SO)
Institute of Informatics
Sierre, Switzerland
henning.mueller@hevs.ch

Weidong Cai
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
tom.cai@sydney.edu.au

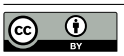
Abstract

Medical visual question answering (Med-VQA) focuses on analyzing medical images to accurately respond to clinicians' specific questions. Although integrating prior knowledge can enhance VQA reasoning, current methods often struggle to extract relevant information from the vast and complex medical knowledge base, thereby limiting the models' ability to learn domain-specific features. To overcome this limitation, our study presents a novel information mining approach that leverages large language models (LLMs) to efficiently retrieve pertinent data. Specifically, we design a latent knowledge generation module that employs LLMs to separately extract and filter information from questions and answers, enhancing the model's inference capabilities. Furthermore, we propose a multi-level prompt fusion module in which an initial prompt interacts with the extracted latent knowledge to draw clinically relevant details from both unimodal and multimodal features. Experimental results demonstrate that our approach outperforms current state-of-the-art models on multiple Med-VQA benchmark datasets.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; **Natural language processing**.

*Also with Department of Radiology and Medical Informatics, The University of Geneva, Switzerland.
†Also with The Sense Innovation and Research Center, Lausanne & Sion, Switzerland.



WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717556

Keywords

Medical Visual Question Answering, Large Language Models, Multimodal Learning, Knowledge Injection

ACM Reference Format:
Ao Ma, Zhiyuan Li, Zhuonan Liang, Tiancheng Gu, Jianan Fan, Jieting Long, Henning Müller, and Weidong Cai. 2025. Seek Inner: LLM-Enhanced Information Mining for Medical Visual Question Answering. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3701716.3717556

1 Introduction

Medical visual question answering (Med-VQA) [39] is an essential branch of multimodal learning [20] in the medical domain, aiming to answer questions related to medical images by integrating visual and linguistic information. Recently, Med-VQA has become increasingly important in healthcare by streamlining workflows, reducing radiologists' workload [45], enhancing patient understanding of their conditions [10], and offering a reliable 'second opinion' in clinical diagnoses to minimize the risk of misdiagnosis [23]. A fundamental Med-VQA model adopts a joint embedding [2] framework with four main components: an image encoder, a text encoder, a feature fusion module, and an answering component. The image encoder often uses CNN architectures [54] [22], while the text encoder processes text sequences with models like LSTM [24] or Transformer [57]. The feature fusion module integrates and captures latent relationships between visual and textual features, often through the attention mechanism [64] and the pooling module [67]. Finally, the answer component produces either classification-based or free text outputs. However, current Med-VQA [1] still faces several critical challenges: (1) **High accuracy requirements:** Med-VQA predictions directly impact treatment decisions, where

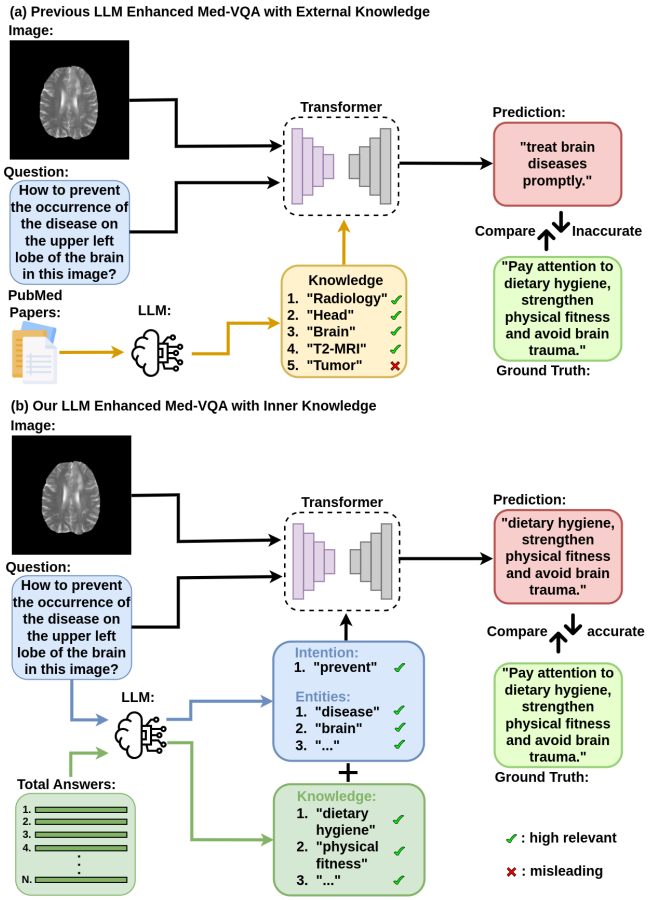


Figure 1: A comparison between our proposed LLM-enhanced Med-VQA with inner knowledge and previous methods.

errors can lead to misdiagnosis and endanger patient safety. (2) **Lack of datasets:** Medical data sensitivity and the need for expert annotation limit the size of the dataset, making high-quality question-answer pair generation challenging. (3) **Imbalanced data distribution:** The dominance of normal images and uneven disease distribution hinder model generalization, affecting prediction accuracy.

To address these challenges, most existing methods rely on large-scale external datasets for the pretraining and subsequent fine-tuning strategy to boost Med-VQA performance [28]. However, this strategy not only demands substantial computational resources but also faces issues of domain specificity. To alleviate the burden of pretraining, some researchers [60] have explored knowledge injection techniques that integrate external knowledge to enhance the model's comprehension and reasoning capabilities. However, external knowledge injection still presents several challenges. First, the injection process may introduce redundant and irrelevant information, increasing computational overhead and reducing inference efficiency [69]. Additionally, the risk of incorporating misleading knowledge may result in erroneous conclusions and undermine the reliability of clinical diagnoses, which may not only impair the reasoning ability of Med-VQA models but also pose potential risks to patient safety.

Recently, LLMs [66] have demonstrated outstanding performance in text understanding and semantic reasoning tasks [34–36], which makes their application in Med-VQA show great potential. LLMs can notably reduce costs and improve task efficiency [43], particularly in tasks with moderate accuracy requirements. In the field of Med-VQA, LLMs reduce costs by automatically generating high-quality question-answer pairs, thereby minimizing reliance on expensive expert annotations.

In this paper, we propose an LLM-enhanced Latent Knowledge Prompt Fusion Model, as shown in Figure 1, aiming to deeply explore the latent knowledge [58] within the data and model itself while avoiding interference caused by irrelevant external knowledge [44]. Compared to previous methods that use LLMs to directly answer questions, they struggled to precisely answer specialized medical questions because their knowledge is primarily based on general common sense rather than domain-specific expertise [18, 19, 38, 56]. Additionally, utilizing LLMs to generate answers often leads to uncertainty and hallucinations [65] [47]. In contrast, our model leverages the LLM for mid-level knowledge extraction and vocabulary refinement [59], rather than using them as the answer generation module. This effectively avoids these shortcomings and ensures more reliable and accurate responses. First, we design a latent knowledge prompt generation module, utilizing LLMs to analyze intention and extract entities within the question, ensuring an accurate understanding of the core of the question. Additionally, we employ the LLM to refine the answer by filtering out irrelevant information to optimize knowledge quality. Subsequently, we propose a multi-level prompt fusion module. In this module, independent prompts are assigned at each level, guiding their fusion with corresponding multimodal features. At each level, the prompts interact with the latent knowledge from both the question and the answer. These prompts are then integrated with text-only features, image-only features, and text-image fusion representations, enabling the model to extract critical clinical knowledge and enhance medical reasoning capabilities. Finally, we aggregate prompts from multiple level into a final prompt, which incorporates rich information from different layers and modalities, contributing to the final answer prediction.

The main contributions of our work can be summarized as follows:

- We propose an LLM-enhanced latent knowledge prompt generation module, which utilizes LLM to extract implicit knowledge from both the question and the answer to generate prompts, effectively avoiding interference caused by the injection of irrelevant external knowledge.
- We propose a multi-level prompt fusion module, where independent prompts are assigned at each level. These prompts, enriched with latent knowledge, are fused with text-only, image-only, and text-image features to extract critical clinical knowledge and enhance medical reasoning capabilities. Finally, multi-level prompts are aggregated into a final prompt to optimize the final answer prediction.
- Our proposed model achieves state-of-the-art performance on SLAKE [41], VQA-RAD [31], and MED-VQA 2019 datasets [3].

2 Related Work

2.1 Medical Visual Question Answering

Medical visual question answering (Med-VQA) [6] is an essential branch of medical image analysis in multimodal learning, aiming to enhance healthcare quality and efficiency as an auxiliary diagnostic tool [17]. However, Med-VQA remains an evolving field. Early approaches [4, 11, 49] adopted transfer learning [21], where models were pre-trained on large-scale natural image datasets [13] and then fine-tuned on target medical datasets. Nevertheless, the effectiveness of transfer learning is limited due to the small size of Med-VQA datasets and the inherent domain gap between natural and medical images [62]. Meta-learning [15] has been proposed as another solution to address data limitations. Nguyen et al. introduced the Multimodal Enhanced Visual Features (MEVF) method [46], which uses meta-learning units and unsupervised autoencoders to train visual encoders. This approach generates structured labels and develops task-specific solutions for VQA-RAD [31]. However, it requires additional labelled data and struggles with unseen domains due to the variability of medical tasks. With the success of CLIP [50] in leveraging large-scale image-text pairs for general domain tasks, more experiments have adopted cross-modal pre-training techniques. Methods like M2I2 [33] and M3AE [7] combine self-supervised learning [27] with masked image modelling and masked language modelling, enabling models to jointly learn [72] representations of medical images and their corresponding descriptions. PubMedCLIP [14] pretrains models using radiology images from the ROCO dataset [48], while BioMedCLIP [70] extends the CLIP architecture by integrating the larger PMC-15M [71] [37] biomedical literature corpus. Although these methods leverage larger datasets for pretraining, their reliance on pre-defined labels limits their ability to comprehend and integrate external knowledge effectively. To address these challenges, we propose a method that injects knowledge prompts into the model to enhance the final answer prediction process. Additionally, we utilize the LLMs for knowledge construction, which simplifies the workflow and improves efficiency.

2.2 Knowledge Injection-based Visual Question Answering

Knowledge-driven visual question answering (VQA) relies on knowledge acquisition and integration. Early methods parsed input queries and retrieved supporting knowledge from fixed knowledge bases (KBs) to generate answers [30], but they performed poorly on complex medical reasoning tasks. In recent years, researchers have focused on multimodal reasoning [42], incorporating knowledge graphs [25] and external medical ontologies to enhance Med-VQA's reasoning capabilities [26]. Meanwhile, large language models (LLMs) such as ChatGPT [61] have been used as implicit knowledge engines, either by directly predicting answers or extracting relevant evidence. However, due to the lack of precise visual information representation, LLMs remain limited in medical reasoning tasks. To address this, researchers have proposed methods such as PICa [63] and Prophet [51], which generate image captions [9] to guide LLM-based reasoning. However, descriptive text generation remains challenging in the medical field, as it requires a high level of domain expertise. Moreover, studies [69] have shown that external knowledge injection is not always effective, as some injected knowledge may be irrelevant or redundant, forcing the model to process excessive information, increasing computational costs, and reducing reasoning efficiency. For instance, it was found that when using medical knowledge bases such as UMLS [5], models often relearn known concepts, leading to inefficient knowledge utilization. Many existing methods fail to precisely align the injected knowledge with the question's actual requirements, resulting in incorrect reasoning paths. Additionally, both medical knowledge bases and LLMs may contain erroneous, outdated, or informal medical information, which can mislead Med-VQA models. For example, [32] discovered that large vision-language models (VLLMs) sometimes generate incorrect medical recommendations based on non-expert data, leading to serious reasoning biases.

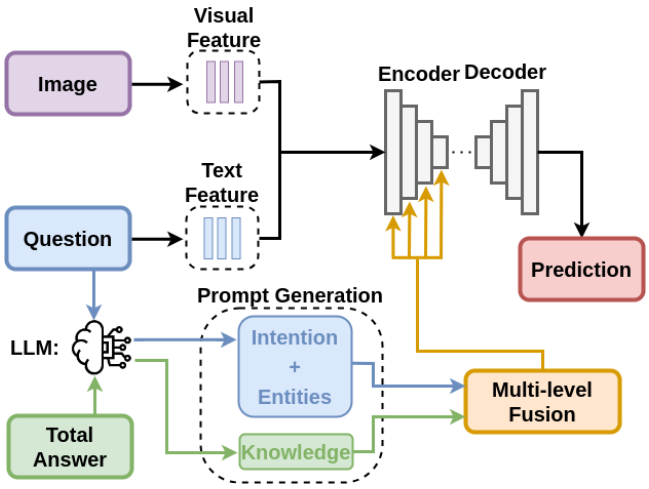


Figure 2: The overview structure of our proposed method

In contrast, our proposed model mines intrinsic data and latent knowledge, reducing interference from irrelevant external knowledge. Furthermore, concise and refined knowledge prompts improve knowledge fusion efficiency, optimizing Med-VQA's reasoning performance.

3 Method

The overall architecture of our proposed LLM-enhanced Latent Knowledge Prompt Fusion Model (LKPF) in Med-VQA is illustrated in Figure 2. The model consists of the following two core modules: the Latent Knowledge Prompt Generation Module (Section 3.1) and the Multi-level Prompt Fusion Module (Section 3.2). This structured design ensures effective integration of medical knowledge, enhancing the model's ability to generate accurate and context-aware responses.

3.1 Latent Knowledge Prompt Generation Module

Our proposed latent knowledge prompt generation module consists of two processes: the latent knowledge extraction process

driven by large language models and the prompt generation process. As shown in Figure 3, we will provide a detailed introduction below. In the process of latent knowledge extraction, we utilize large language models to extract knowledge from both questions and answers separately. For the question part, we first employ a large language model to identify the intention of the question and map it to a predefined list of question intentions. Then, we extract key entities from the question, forming one or more entities based on the question. The process of question intention recognition and key entity extraction can be formulated as follows:

$$I_Q = f_{LLM_I}(Q), \quad I_Q \in \mathcal{I}, \tag{1}$$

$$\mathcal{E}(Q) = f_{LLM_E}(Q), \tag{2}$$

where Q represents the input question, $\mathcal{I} = \{I_1, I_2, ..., I_{11}\}$ is the predefined set of question intention categories (containing 11 categories such as abnormality, plane, position, modality, organ, relationship, size, pervent, pathology, structure, and other). f_{LLM_I} is the intention recognition function of the large language model, which maps the question Q to an intention category I_Q. $\mathcal{E}(Q)$ represents the set of entities extracted from the question Q, generated by the entity extraction function f_{LLM_E}.

For the answer part, we first determine the image classification associated with the answer to narrow down the answer scope. Then, within this restricted scope, we use the large language model to filter the answers, retaining only key information, thereby refining the knowledge within the answer scope. The process of answer classification and knowledge extraction can be formulated as follows:

$$C_A = f_{cls}(A), \quad C_A \in \mathcal{C}, \tag{3}$$

$$\mathcal{K}(A) = f_{LLM_F}(A, C_A), \tag{4}$$

where A represents the answer corresponding to question Q. $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ is the predefined set of answer categories (based on image classification, they are abnormality, modality, organ, position, and plane). f_{cls} is the classification function, which maps the answer A to its corresponding category C_A. f_{LLM_F} is the keyword extraction and filtering function, responsible for refining the answer content. C_A is the category to which the answer A belongs. $f_{LLM_F}(A, C_A)$ represents the filtering process within the constraints of category C_A, ensuring that only key information is retained. $\mathcal{K}(A)$ is the final extracted knowledge set from answer A.

In the prompt generation module, we encode the latent knowledge derived from the question intention, question entities, and answer knowledge using a text encoder, obtaining the following three features:

$$F_{in} = E_T(I_Q), \tag{5}$$

$$F_e = E_T(\mathcal{E}(Q)), \tag{6}$$

$$F_k = E_T(\mathcal{K}(A)), \tag{7}$$

where F_{in} represents the question intention feature. F_e represents the question entity feature. F_k represents the answer knowledge feature. $E_{T(\cdot)}$ represents the text encoder. These three features are the key to prompt generation.

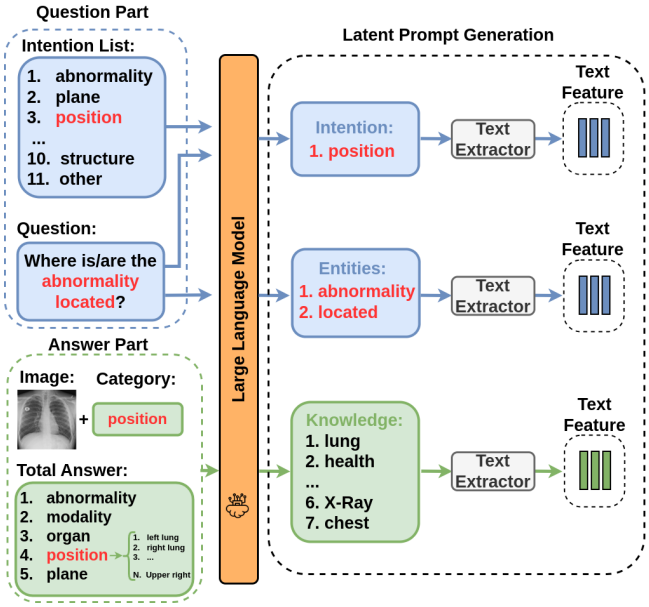


Figure 3: The process of latent knowledge prompt generation module to utilize LLMs to extract latent knowledge from questions and answers, then generate latent prompts.

3.2 Multi-level Prompt Fusion Module

In this section, we introduce the Multi-Level Prompt Fusion Module and explain its implementation in detail. We first concatenate the three features (F_{in}, F_e, F_k) obtained from the Latent Knowledge Prompt Generation Module to obtain our prompt feature X_P, as shown in the following equation:

$$X_P = Concat(F_{in}, F_e, F_k), \tag{8}$$

where $Concat(\cdot)$ denotes the concatenation operation for feature vectors.

After obtaining the prompt vector, we proceed with multi-level prompt fusion to refine the prompt representation from both depth and breadth. The detailed process is illustrated in Figure 4. We construct a multi-level prompt fusion model to fully capture semantic information across different levels. First, we use CLIP-ViT [52] and RoBERTa [12] to extract image and language features, respectively. Then, self-attention is applied to obtain unimodal features F_I and F_T as follows:

$$F_I = SA(E_I(X_I)), \tag{9}$$

$$F_T = SA(E_T(X_T)), \tag{10}$$

where X_I and X_T represent the input image data and text data, respectively. $E_I(\cdot)$ and $E_T(\cdot)$ are the corresponding image encoder and text encoder. $SA(\cdot)$ denotes the self-attention mechanism used for extracting unimodal features.

After obtaining the unimodal representations, we employ cross-attention to fuse the image and language features, yielding the multimodal feature representation:

$$F_M = Concat\left[Proj\left(CA(F_I, F_T, F_T)\right), Proj\left(CA(F_L, F_I, F_I)\right)\right], \tag{11}$$

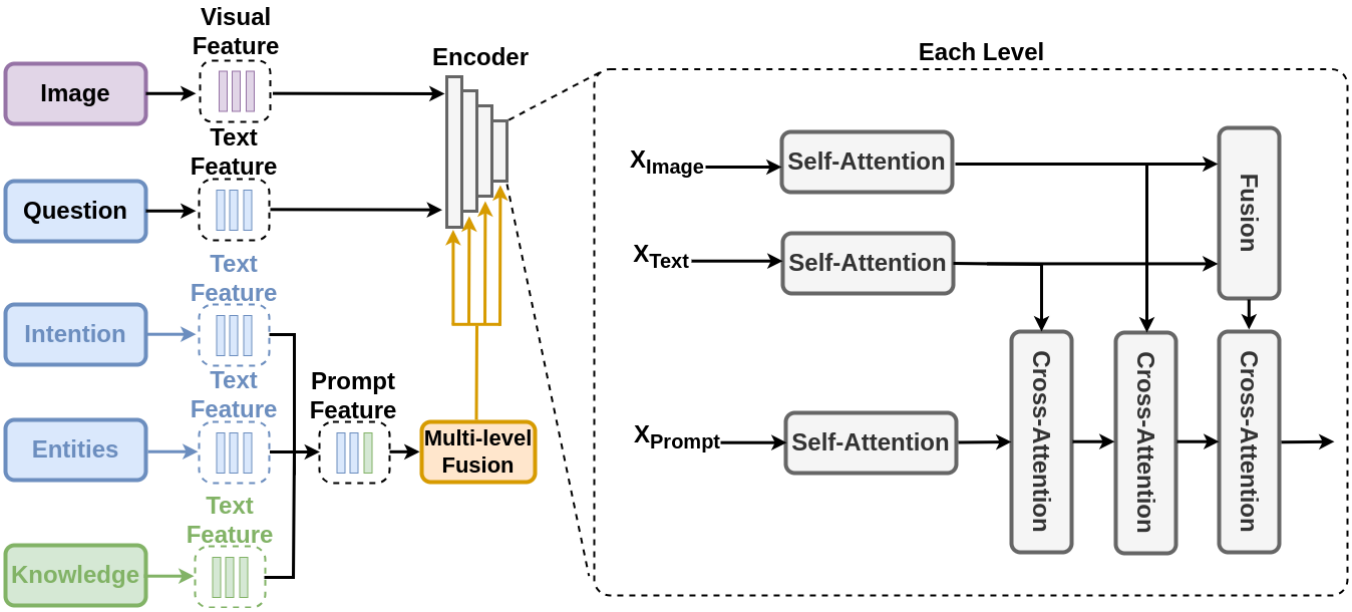


Figure 4: Framework of our proposed Multi-level Prompt Fusion Model. In this module, features from different modalities are efficiently fused to generate a latent prompt.

where $CA(\cdot)$ represents the cross-attention mechanism, which facilitates multimodal fusion. $Proj(\cdot)$ denotes the projection layer, which is responsible for aligning and refining the feature representations. After obtaining the unimodal features F_I, F_T, and the multimodal features F_M, we introduce the latent prompt fusion module to further integrate clinic-relevant information through cross-attention.

We utilize the prompt to progressively integrate with textual, visual, and multimodal information. The specific steps are as follows:

$$X_{Pt} = CA(X_P, F_T, F_T), \tag{12}$$

$$X_{Pti} = CA(X_{Pt}, F_I, F_I), \tag{13}$$

$$X_{Pm} = CA(X_{Pti}, F_M, F_M). \tag{14}$$

After demonstrating a complete workflow of a prompt feature, we take a macroscopic view of the multi-level prompt module. In this module, each layer's prompt is independent and learnable, ensuring the model's scalability and flexibility. ℓ layer's prompt vector X^{ℓ}_{Pm} consists of a length P and a hidden dimension d, and it undergoes a pooling operation to obtain the level prompt representation v^{ℓ}. Then, the pooled vectors from all layers are stacked into a three-dimensional tensor V for subsequent inter-layer interactions:

$$X^{\ell}_{Pm} \in \mathbb{R}^{P\times d}, \quad v^{\ell} = \text{Pool}(X^{\ell}_{Pm}) \in \mathbb{R}^{B\times d}, \tag{15}$$

$$V = \{v^1, v^2, \ldots, v^L\} \in \mathbb{R}^{B\times L\times d}, \tag{16}$$

where B is the batch size, L is the number of multi-level prompt layers, and $Pool(\cdot)$ represents the pooling operation, such as average pooling or max pooling.

To capture inter-layer dependencies, a multi-head self-attention mechanism is employed, which enhances the global prompt representation by modelling the interactions among different layers:

$$C = MultiHeadAttention(V, V, V) \in \mathbb{R}^{B\times L\times d}, \tag{17}$$

where C represents the attention-enhanced multi-level prompt representation. Finally, the globally pooled prompt vector is obtained through average pooling:

$$X_{Pfinal} = \frac{1}{L}\sum_{\ell=1}^{L} C^{\ell} \in \mathbb{R}^{B\times d}. \tag{18}$$

Finally, we combine the final prompt features with the multimodal information and feed them into downstream tasks for prediction.

4 Experiments

4.1 Datasets

We conducted experiments on three widely used Med-VQA benchmark datasets to evaluate the effectiveness of our proposed method. A brief description of each dataset is provided below.

VQA-RAD [31] contains 315 radiology images and 3,064 question-answer (QA) pairs, with 2,613 QA pairs used for training and 451 QA pairs for testing. The images are evenly distributed across the head, chest, and abdomen. Each image is associated with multiple questions that cover various aspects of radiological interpretation.

SLAKE [41] consists of 642 radiology images and 7,014 QA pairs, encompassing multiple medical imaging modalities and different

Table 1: Comparison of our model with state-of-the-art methods on VQA-RAD, SLAKE, and MED-VQA 2019 datasets.

Methods	VQA-RAD			SLAKE			MED-VQA 2019
	Open	Closed	**Overall**	Open	Closed	**Overall**	**Overall**
SAN [64]	31.30	69.50	54.30	74.00	79.10	76.00	-
BAN [29]	37.40	72.10	58.30	74.60	79.10	76.50	-
MEVF+SAN [26]	49.20	73.90	64.10	75.30	78.40	76.50	68.90
MEVF+BAN [46]	49.20	77.20	66.10	77.80	79.80	78.60	77.86
CPRD+BAN [40]	52.50	77.90	67.80	79.50	83.40	81.10	-
PubMedCLIP [14]	60.10	80.00	72.10	78.40	82.50	80.10	-
MMBERT [28]	63.10	77.90	72.00	-	-	-	67.20
UNICLAM [68]	59.80	82.60	73.20	81.10	85.70	83.10	-
M2I2 [33]	61.80	81.60	73.70	74.70	**91.10**	80.20	-
MITER [53]	59.40	80.50	72.10	79.20	84.40	81.20	-
ARL [8]	65.10	**85.96**	77.55	79.70	89.30	84.10	79.80
M3AE [7]	67.23	83.46	77.01	80.31	87.82	83.25	79.87
Ours	**71.53**	83.19	**79.20**	**83.18**	87.74	**84.92**	**81.31**

anatomical regions. All images are annotated by experienced physicians. The dataset was divided into a ratio of 70:15:15 for training, validation, and testing, respectively.

MED-VQA 2019 [3] comprises 3,200 radiology images. It features four primary categories of questions: Modality, Plane, Organ System, and Abnormality. These categories are designed with varying degrees of difficulty, incorporating text generation approaches.

In the VQA-RAD and SLAKE datasets, questions are categorized into open-ended (free-text responses) and closed-ended (e.g., Yes/No) formats. This experimental evaluation provides a comprehensive assessment of our method's performance across different medical VQA tasks.

4.2 Implementation Details

Baseline We adopt M3AE as our baseline model. M3AE is designed to map medical images and text into a joint space for visual and language representation learning in the medical domain. It achieves this by reconstructing pixels and tokens from randomly masked images and texts, which is proven to be well performed in the report generation domain [16]. In our implementation, we utilize its pre-trained weights from the ROCO [48] and MedICaT [55] datasets. **Settings** Our model consists of CLIP-ViT-B [52] as the image feature extractor and RoBERTa [12] as the language feature extractor. The number of multimodal fusion blocks and multi-level prompt modules is set to 6. The training is conducted on a single NVIDIA GeForce RTX 3090 GPU using the AdamW optimizer with a learning rate of 5e-6. Input images are resized to 384 × 384, and the feature dimension is set to 768.

4.3 Comparison Experiments

We evaluated our model against state-of-the-art methods on the VQA-RAD, SLAKE, and MED-VQA 2019 datasets, using accuracy as our evaluation metric in line with previous studies. Detailed comparisons are provided in Table 1. The experimental results reveal that our model consistently outperforms existing techniques across all datasets. On VQA-RAD, our model achieves an overall accuracy of 79.20%, with a remarkable 71.53% in the open-ended category and a competitive 83.19% in the closed-ended category when compared to ARL's 85.96%. For the SLAKE dataset, the model excels with an overall accuracy of 84.92%, notably reaching 87.74% in the closed category. Additionally, on MED-VQA 2019, our approach attains an accuracy of 81.31%, surpassing ARL's 79.80% and M3AE's 79.87%, thereby demonstrating robust generalization capabilities. In summary, our model not only achieves superior performance overall but also exhibits outstanding results in open-ended questions, underscoring its potential value for real-world Med-VQA applications.

Table 2: The ablation study for the LKPF model was conducted on the SLAKE to verify the contribution of LKPG and MPF modules to the overall performance.

Method	Open	Closed	**Overall**
Baseline (M3AE)	80.31	87.82	83.25
+LKPG (A-only)	80.93	**88.46**	83.88
+LKPG (Q-only)	81.40	88.22	84.07
+LKPG	81.71	88.22	84.26
+MPF	80.71	87.98	83.60
+MPF & LKPG	**83.18**	87.74	**84.92**

4.4 Ablation Study

In this subsection, we conduct an ablation study to verify the effectiveness of the proposed large language model-driven latent knowledge prompt generation module (LKPG) and multi-level prompt fusion module (MPF). LKPG-A represents that we only consider the latent knowledge from the answers, while LKPG-Q is based on

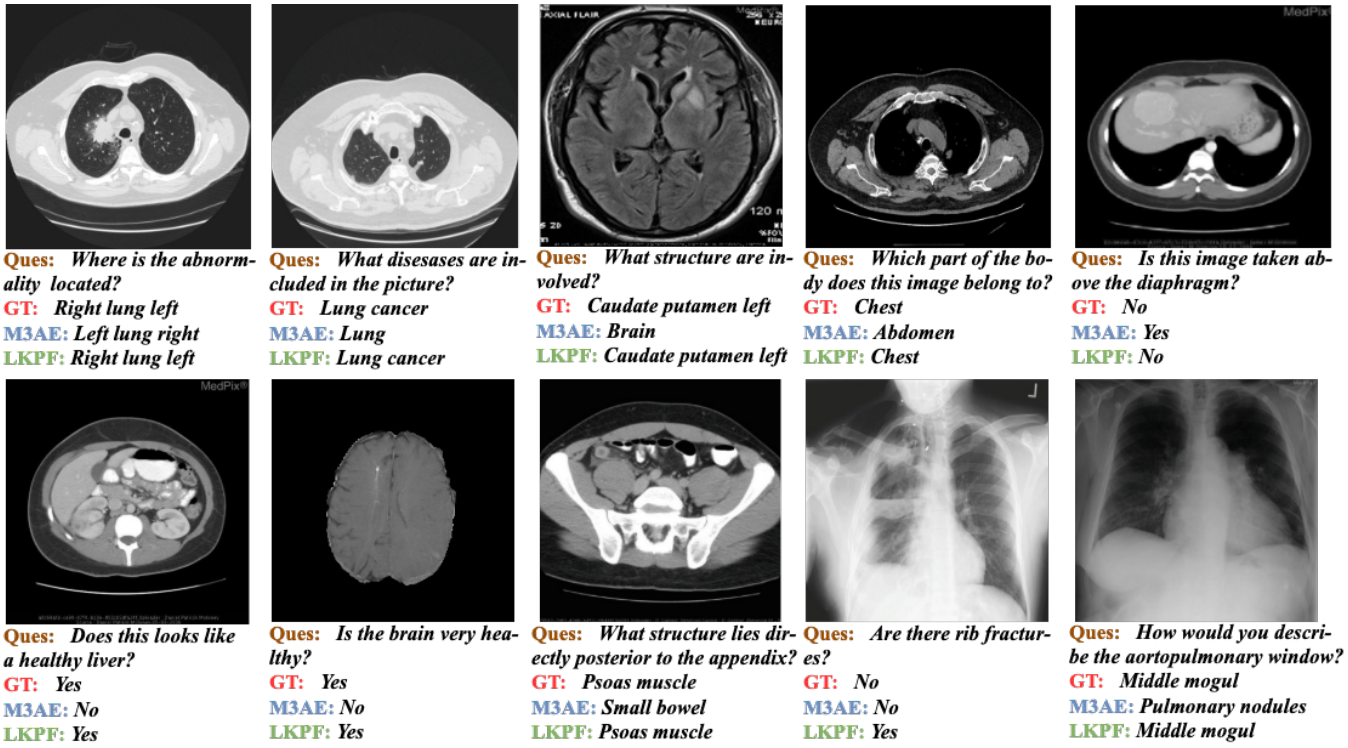


Figure 5: Results of the cases for the baseline model (M3AE) and the proposed model (LKPF) from the datasets. Red and green denote the wrong and correct predictions, respectively.

questions. We separately examined the independent effects of the LKPG and MPF components, as shown in Table 2.

Introducing the LKPG component alone improved the overall accuracy of the SLAKE dataset from 83.25% to 84.26%. To further analyze its contribution, we decomposed the LKPG module into two parts: knowledge based on questions (LKPG-Q) and knowledge based on answers (LKPG-A). Experimental results indicate that both components enhance performance, with LKPG-Q yielding a greater improvement (84.07%) compared to LKPG-A (83.88%), suggesting that question-related knowledge has a more important impact on the model.

Introducing the MPF component alone led to a slight improvement (83.60%), but the effect was not substantial. This is likely because MPF does not carry crucial semantic information when used independently, limiting its impact on model enhancement.

Finally, when integrating both LKPG and MPF into the baseline model, we observed a more marked performance boost, with accuracy increasing from 83.25% to 84.92%. This indicates that the combination of both components provides complementary advantages, further enhancing the model's overall performance.

4.5 Qualitative Analysis

In this subsection, we present an in-depth qualitative comparison between our approach and the baseline M3AE, highlighting the differences in performance through detailed visualizations presented in Figure 5. The figure illustrates side-by-side results on multiple datasets, providing clear evidence of the improvements achieved by our method.

Our analysis covers various cases involving both open-ended and closed-ended questions, and it demonstrates that our latent knowledge prompt fusion method substantially enhances the model's ability to extract and utilize relevant information. This enhanced capability leads to more accurate answers in benchmark evaluations. Notably, in all three open-ended questions examined, our model consistently produced precise responses. For example, in the first case, the baseline model reversed the correct answer, leading to a clear misinterpretation, whereas our model was able to successfully identify and provide the correct response. Additionally, in the third case, the question specifically probed the internal structure of the brain. Here, our method effectively captured the intended focus on "structure," enabling it to correctly answer with "caudate putamen left parietal." In contrast, the baseline misinterpreted the query as referring solely to an organ, which resulted in a more generic and inaccurate answer, "brain."

These detailed case studies not only underscore the superior performance of our model in handling intricate queries but also highlight the robustness and versatility of our latent knowledge prompt fusion technique. Overall, the qualitative evidence strongly supports our claim that the proposed method substantially improves the model's performance across both open-ended and closed-ended Med-VQA tasks, demonstrating its potential for practical, real-world applications in medical visual question answering.

5 Conclusion

This study proposes an LLM-enhanced latent knowledge prompt fusion model, aiming to improve performance in Med-VQA tasks. The model leverages its inner latent knowledge to understand questions and improve its reasoning capabilities.

We introduce two key innovative modules. The first is the latent knowledge prompt generation module, which utilizes LLMs to extract the intention and key entities from the question. By incorporating refined answer information, it generates latent prompts that help the model be aware of its own knowledge, thereby avoiding irrelevant or misleading external knowledge interference. Additionally, we design a multi-level prompt fusion model, which employs multiple independent prompt levels. This design is both flexible and scalable, enabling the effective utilization of semantic information at different levels. At each level, the prompts interact with multiple modalities, and ultimately, all prompts are aggregated into a final prompt, contributing to the final answer prediction.

Empirical validation on three authoritative benchmark datasets demonstrates that our approach achieves outstanding answer prediction performance in Med-VQA tasks. Looking ahead, we plan to deploy the latent knowledge prompt fusion model into complex answer generation models to facilitate efficient and accurate reasoning.

References

[1] Suheer Al-Hadhrami, Mohamed El Bachir Menai, Saad Al-Ahmadi, and Ahmed Alnafessah. 2023. A critical analysis of benchmarks, techniques, and models in medical visual question answering. *IEEE Access* 11 (2023), 136507–136540.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

[4] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

[5] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[6] Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, et al. 2024. A survey of medical vision-and-language applications and their techniques. *arXiv preprint arXiv:2411.12195* (2024).

[7] 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 (2024), 103018.

[8] Zhihong Chen, Guanbin Li, and Xiang Wan. 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5152–5161.

[9] Fuze Cong, Shibiao Xu, Li Guo, and Yinbing Tian. 2022. Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3569–3577.

[10] Nathan M Cross, Joseph Wildenberg, Geraldine Liao, Sean Novak, Thomas Bevilacqua, James Chen, Evan Siegelman, and Tessa S Cook. 2020. The voice of the radiologist: enabling patients to speak directly to radiologists. *Clinical imaging* 61 (2020), 84–89.

[11] Alba Garcia Seco De Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K Antani, and Henning Müller. 2013. Overview of the ImageCLEF 2013 medical tasks. *CLEF (Working Notes)* 1179 (2013).

[12] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3255–3265.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[14] 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*. 1181–1193.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[16] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2024. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7995–8004.

[17] Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. 2024. LaPA: Latent Prompt Assist Model For Medical Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[18] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. (2022), 956–968.

[19] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10867–10877.

[20] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *Ieee Access* 7 (2019), 63373–63394.

[21] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. 2018. Overview of imageclef 2018 medical domain visual question answering task. *Proceedings of CLEF 2018 Working Notes* (2018).

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[23] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).

[24] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).

[25] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4156–4165.

[26] Jian Huang, Yihao Chen, Yong Li, Zhenguo Yang, Xuehao Gong, Fu Lee Wang, Xiaohong Xu, and Wenyin Liu. 2023. Medical knowledge-based network for patient-oriented visual question answering. *Information Processing & Management* 60, 2 (2023), 103241.

[27] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[28] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1033–1036.

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

[30] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 11 (2022), 11196–11215.

[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 (2018), 1–10.

[32] 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 (2024).

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

[34] Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. 2024. Llms for relational reasoning: How far are we?. In *Proceedings of the 1st International Workshop on Large Language Models for Code*. 119–126.

[35] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. 2024. Enhancing Advanced Visual Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 1915–1929.

[36] Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. 2024. Multimodal Causal Reasoning Benchmark: Challenging Vision Large Language Models to Infer Causal Links Between Siamese Images. *arXiv preprint arXiv:2408.08105* (2024).

[37] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training

using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 525–536.

[38] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems* 35 (2022), 10560–10571.

[39] 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* 143 (2023), 102611.

[40] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021. 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*. Springer, 210–220.

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

[42] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[43] Wenhan Lyu, Yimeng Wang, Tingting Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 63–74.

[44] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[45] Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes. 2015. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology* 22, 9 (2015), 1191–1198.

[46] 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*. Springer, 522–530.

[47] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[48] 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*. Springer, 180–189.

[49] Yalei Peng, Feifan Liu, and Max P Rosen. 2018. UMass at ImageCLEF Medical Visual Question Answering (Med-VQA) 2018 Task.. In *CLEF (working notes)*. 1–9.

[50] 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*. PMLR, 8748–8763.

[51] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 14974–14983.

[52] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutz. 2022. How Much Can CLIP Benefit Vision-and-Language Tasks?. In *ICLR*.

[53] 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 (2024), 121526.

[54] K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

[55] 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* (2020).

[56] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 951–967.

[57] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[58] John R Vokey and Philip A Higham. 1999. Implicit knowledge as automatic, latent knowledge. *Behavioral and Brain Sciences* 22, 5 (1999), 787–788.

[59] Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, and Yuxiang Li. 2024. An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study. *Journal of Medical Internet Research* 26 (2024), e54580.

[60] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1290–1296.

[61] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.

[62] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis* 69 (2021), 101985.

[63] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 3081–3089.

[64] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[65] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469* (2023).

[66] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.

[67] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 1821–1830.

[68] Chenlu Zhan, Peng Peng, Hongwei Wang, Gaoang Wang, Yu Lin, Tao Chen, and Hongsen Wang. 2025. Uniclam: Contrastive representation learning with adversarial masking for unified and interpretable medical vision question answering. *Medical Image Analysis* (2025), 103464.

[69] Ningyu Zhang, Shumin Deng, Xu Cheng, Xi Chen, Yichi Zhang, Wei Zhang, Huajun Chen, and Hangzhou Innovation Center. 2021. Drop redundant, shrink irrelevant: Selective knowledge injection for language pretraining.. In *IJCAI*. 4007–4014.

[70] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).

[71] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* (2023).

[72] Yuan Zhou, Jing Mei, Yiqin Yu, and Tanveer Syeda-Mahmood. 2023. Medical visual question answering using joint self-supervised learning. *arXiv preprint arXiv:2302.13069* (2023).