Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering

Wang, Jianfeng; Seng, Kah Phooi; Shen, Yi; et.al.

Wang, J., Seng, K. P., Shen, Y., Ang, L.-M., & Huang, D. (2024). Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering. Electronics, 13(12), 1–12. https://doi.org/10.3390/electronics13122273
Document Type: Published Version

Link to Published Version: https://doi.org/10.3390/electronics13122273

electronics

MDPI

Article

Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering

Jianfeng Wang [1], Kah Phooi Seng [1,2,3,*], Yi Shen [4], Li-Minn Ang [3] and Difeng Huang [5]

[1] XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong Liverpool University, Suzhou 215000, China; jianfeng.wang17@student.xjtlu.edu.cn
[2] School of Computer Science, Queensland University of Technology, Brisbane City, QLD 4000, Australia
[3] School of Science, Technology and Engineering, University of the Sunshine Coast, Petrie, QLD 4502, Australia; lang@usc.edu.au
[4] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215011, China; sy@post.usts.edu.cn
[5] School of Mathematics and Economics, Hubei University of Education, Wuhan 430062, China; dfeng20@163.com
* Correspondence: jasmine.seng@xjtlu.edu.cn

Abstract: Medical Visual Question Answering (Med-VQA) faces significant limitations in application development due to sparse and challenging data acquisition. Existing approaches focus on multi-modal learning to equip models with medical image inference and natural language understanding, but this worsens data scarcity in Med-VQA, hindering clinical application and advancement. This paper proposes the ITLTA framework for Med-VQA, designed based on field requirements. ITLTA combines multi-label learning of medical images with the language understanding and reasoning capabilities of large language models (LLMs) to achieve zero-shot learning, meeting natural language module needs without end-to-end training. This approach reduces deployment costs and training data requirements, allowing LLMs to function as flexible, plug-and-play modules. To enhance multi-label classification accuracy, the framework uses external medical image data for pretraining, integrated with a joint feature and label attention mechanism. This configuration ensures robust performance and applicability, even with limited data. Additionally, the framework clarifies the decision-making process for visual labels and question prompts, enhancing the interpretability of Med-VQA. Validated on the VQA-Med 2019 dataset, our method demonstrates superior effectiveness compared to existing methods, confirming its outstanding performance for enhanced clinical applications.

Keywords: medical visual question answering (Med-VQA); large language models (LLMs); multi-label learning; attention mechanisms; zero-shot learning

Citation: Wang, J.; Seng, K.P.; Shen, Y.; Ang, L.-M.; Huang, D. Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering. *Electronics* **2024**, *13*, 2273. https://doi.org/10.3390/electronics13122273

Academic Editor: Chang Wook Ahn

Received: 30 April 2024
Revised: 29 May 2024
Accepted: 3 June 2024
Published: 10 June 2024

1. Introduction

Visual Question Answering (VQA) is an interdisciplinary task combining computer vision and natural language processing, where models utilize relevant documents, images, and questions to predict answers to given questions. VQA has been successfully applied in various fields such as search engines and chatbots [1]. However, in specialized domains like biomedicine, VQA systems are rarely used due to numerous constraints [2]. With the rapid development of digital medicine and artificial intelligence technologies, Medical Visual Question Answering (Med-VQA) has emerged as a cutting-edge topic in the fields of computer vision and medical research. Med-VQA aims to automatically answer clinical questions related to medical images by analyzing these images in conjunction with natural language processing techniques. Its goal is to enable computers to construct a multi-modal intelligent system oriented towards medical information, learning from medical images and texts to address clinical diagnostic queries. At a time when medical resources and

expertise are scarce, the clinical application of Med-VQA systems can aid in the assessment and interpretation of clinical diagnoses, playing a crucial role in the advancement of clinical diagnostics [3]. Despite its significant potential, the practical application of Med-VQA faces many challenges, notably the scarcity and difficulty in accessing high-quality medical data, which severely impedes the development of Med-VQA in clinical settings [4].

Current research in Med-VQA predominantly focuses on leveraging deep learning technologies to integrate medical images with associated textual data [5]. This multi-modal approach is designed to equip computer models with the ability to concurrently interpret visual information from images and process related textual data. However, achieving high accuracy with this method generally necessitates extensive, precisely annotated training datasets, which presents significant challenges in the medical field. These challenges stem from the high costs and complex ethical and privacy concerns involved in collecting medical-grade annotated data. Furthermore, existing Med-VQA systems frequently rely on end-to-end training strategies, resulting in models with limited interpretability, a critical drawback in clinical settings. Consequently, there is a pressing need for innovative solutions that effectively address data scarcity and can be swiftly implemented in clinical environments while maintaining transparency in the model's decision-making processes.

The Image to Label to Answer (ITLTA) framework proposed in this paper is specifically designed to address the aforementioned challenges. ITLTA significantly reduces its reliance on large-scale Med-VQA data by decomposing the Med-VQA task into two stages: multi-label learning for images and label-based question answering. This is achieved through a pretraining strategy that utilizes easily accessible medical image data. To enhance the model's capability in extracting features from medical images and improve the performance of feature extraction algorithms on specialized medical data, we initially utilize more easily accessible medical image data for multitask learning pretraining. This is particularly critical when applying Med-VQA in scenarios with scarce data. Subsequently, we develop a joint feature and label attention mechanism that performs cross-modal correlation learning between image feature information and label embeddings, integrating multi-regional image features deeply into label embeddings. During this process, the module also conducts semantic co-occurrence learning among labels, thus facilitating the multi-label classification of medical images in terms of modality, organs, and abnormalities. Furthermore, the framework leverages the powerful language understanding and processing capabilities of large language models to comprehend and respond to medical natural language questions through zero-shot learning, thereby circumventing the need for complex end-to-end model training. Empirical validation on the VQA-Med 2019 dataset [6] has demonstrated excellent results, confirming the effectiveness of our approach.

This study not only advances the development of Med-VQA technology but also holds significant importance and value in promoting its application in actual clinical settings. We hope that this research can provide a new perspective and framework for future applications of Med-VQA.

2. Related Work

In 2018, Hasan and colleagues systematically introduced their research on Med-VQA at the CLEF conference, presenting the foundational dataset VQA-Med2018 and establishing guidelines for related research and data [7]. The structure of Med-VQA models is similar to that of general-domain visual question answering, involving the extraction of multi-modal features, feature fusion, and ultimately, answer prediction [8]. These processes are based on known classifications of questions within the medical QA dataset, where the content of questions must be directly related to the information in the images, and closed-domain questions do not rely on unrelated external information for inference. Building on existing visual question-answering models, subsequent research has been tailored to medical data. For instance, Peng and others utilized the ResNet-152 for transfer learning to extract image features and fine-tuned word vectors for medical information, then employed the attention-based MFB algorithm for feature fusion and a Seq2Seq model for decoding to retrieve

the most relevant answers from the original answer space [9]. Further studies have also adopted a framework that separately extracts features and performs fusion classification before predicting answers [5]. These preliminary efforts have bridged the gap from general visual question answering to Med-VQA, though significant differences remain.

Due to the complexity of medical data features and the challenges associated with labeling, traditional optimization methods in visual question-answering models often struggle to make significant advancements in related research. Therefore, many researchers have proposed various model optimization techniques tailored to the characteristics of medical data. For instance, BD Nguyen and colleagues introduced a framework that employs denoising autoencoders to train on both image and textual features, complemented by few-shot learning techniques, allowing the model to effectively extract features with minimal annotated data [2]. Deepak and others proposed a hierarchical multi-modal network that first analyzes the questioner's queries using a text classification algorithm, then utilizes semantic information derived from this analysis with different types of attention mechanisms to predict answers, enhancing performance in handling complex queries [3]. These efforts are specifically designed to optimize Med-VQA tasks and scenarios. Although there have been improvements in performance, the variability across different datasets prevents these models from meeting the clinical application needs of Med-VQA. As research progresses, Binh D. Nguyen and his team have introduced meta-learning, incorporating noise perturbation to enhance image feature learning [2]. Liming Zhan and others have divided questions into closed and open categories, training inference networks for each category, which resulted in improved reasoning capabilities [10]. These modifications optimize the models from both image and question perspectives to enhance the effectiveness of medical data answering. Still limited by the quantity of existing data and lacking considerations for pretraining with additional data, Haifan Gong and his team categorized images by body parts and pretrained a ResNet network on an external medical dataset, employing multiple attention feature fusion strategies for better feature integration [11]. Haiwei Pan and his team used a multi-perspective attention mechanism with a composite loss algorithm to fuse features, thereby enhancing question analysis and increasing the accuracy of answers [12]. These studies not only incorporated pretraining but also utilized attention mechanisms, significantly improving the outcomes [13]. These efforts to merge different modalities aim to solve visual question-answering challenges, yet the substantial differences between modal features still impede further progress.

Recent large visual language models have achieved remarkable success in various multi-modal tasks [14]. There has also been some research in Med-VQA using large-scale language–vision models from a multi-modal learning perspective [15–17]. These methods have been highly successful, yet their lack of flexibility and the high cost of training impede their practical application in clinical settings. With the tremendous success of large language models like GPT in the field of natural language processing, some studies have also applied large language models to visual question-answering tasks, making significant progress [18,19]. In Med-VQA, although there have been attempts to integrate large language models, the massive corpus data requirements and the complexity of training such models have limited their success [19]. Our study leverages the unique capabilities of large language models for text understanding in the context of Med-VQA, implementing zero-shot learning to meet the natural language processing needs of this field. This approach avoids the need for end-to-end training of large models, significantly reducing deployment costs and the need for extensive training data, while allowing the large language models to serve as flexible, plug-and-play modules. By integrating pretraining with external medical image data, we enhance the model's capability to understand medical image features, thus enabling effective clinical application even in scenarios with limited data.

3. Methods

The Image to Label to Answer (ITLTA) framework is specifically designed for Med-VQA, aimed at addressing issues related to data scarcity and the complexity of model

training inherent in traditional models. The ITLTA framework adopts a phased strategy, dividing the problem-solving process into two main parts: multi-label learning of images and question answering based on these labels with large language models. This approach effectively simplifies the modeling complexity of Med-VQA while maintaining flexibility, thus enhancing the performance of medical question-answering systems.

Initially, we designed a multi-label learning component for medical images tailored to Med-VQA, which involves pretraining on external medical images based on multitask learning. Building on this pretrained model, we then developed a joint image feature and multi-label attention module. This module is capable of performing cross-modal correlation learning between image feature information and label embeddings, as well as semantic co-occurrence learning among labels, thereby identifying and providing multiple labels such as the modality, organ, and abnormalities of medical images. The purpose of this step is to extract a detailed and accurate set of labels from visual data, laying the groundwork for subsequent language processing.

Subsequently, the generated label set, serving as rich visual information, is combined with the question, prompt phrases, and example cases to be input into a large language model. This module utilizes advanced large language models to analyze and respond to the posed questions. By integrating visual and textual information, the module is capable of generating answers corresponding to the questions. Since our visual information is presented in the form of a label set, and the input details for the large model are specific and known, we have a deeper understanding of the entire question-answering decision-making process. This allows us to clearly identify whether incorrect responses are due to insufficient or erroneous visual information, or due to issues related to language comprehension. The structure of this framework is illustrated in Figure 1.

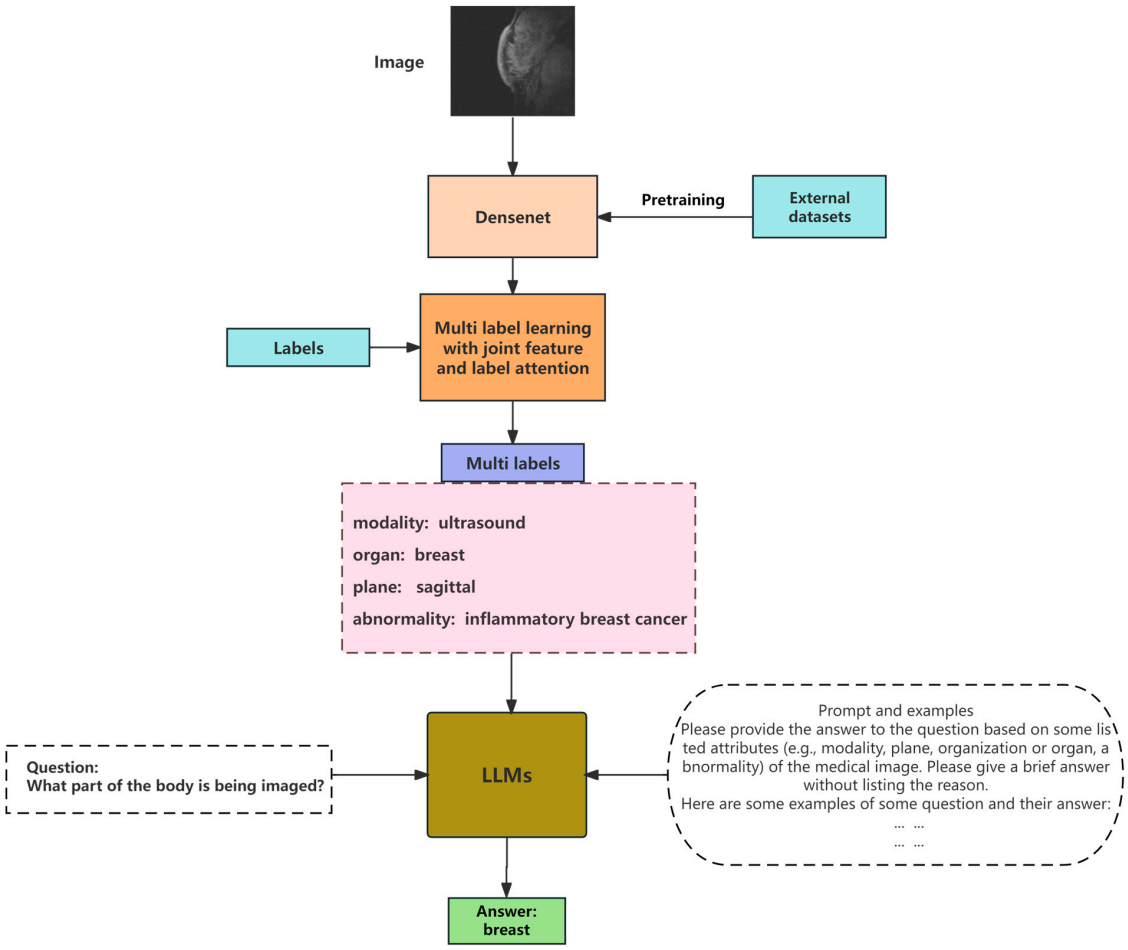

Figure 1. The structure of ITLTA.

3.1. Pretraining Based on Multitask Learning

To enhance the model's capability to extract features from medical images, we utilized expandable additional medical images for multitask learning pretraining. The accuracy of image labels directly affects the final outcomes in Medical Visual Question Answering (Med-VQA). In deep learning-based image recognition tasks, the volume of training data significantly influences the results. Particularly in fields like healthcare, where data are often scarce, the sufficiency of training data crucially determines the quality of the outcomes. Therefore, in these areas, transfer learning often serves as an effective solution to the problem of insufficient training data. Moreover, the more similar the transfer learning data are to the training data, the better the results will be. Thus, in this phase, to achieve desirable results in medical image label learning, we can select other available medical image data based on the data characteristics of our specific application scenario. If the goal is to obtain the best possible results, then initially, it is crucial to choose external pretraining data that are similar to the target task. Subsequently, it is advantageous to use a large amount of external image data. We know that acquiring medical images is much easier than obtaining Med-VQA data, and the volume of publicly available medical images is also significantly larger than that of Med-VQA data. In this study, based on our data scenario, we have chosen publicly available medical images of different modalities and organs for pretraining to mitigate the issue of data scarcity in Med-VQA.

In this study, we created a dataset comprising over 140,000 medical images from five publicly available medical image datasets. These images include multiple common modalities such as MRIs, X-rays, CT scans, and ultrasounds, and feature various organs including the head, chest, limbs, and kidneys, as well as certain diseases. We employed Densenet [20] as the backbone network and designed it to handle multiple learning tasks such as modality classification, organ classification, and disease classification. Since each task involves classifying different attributes of images and these tasks share underlying data, multitask learning with hard parameter sharing is particularly suitable for this research. This approach allows the development of a generic model capable of representing multi-dimensional attributes of medical images. Another benefit of this multitask training approach is that the feature learning for each attribute can mutually enhance each other based on the shared data. In the multitask learning setup, the loss function is designed to simultaneously address any number of classification tasks, with the total loss L defined as the weighted sum of losses from all tasks, as shown below:

$$L = \sum_{i=1}^{N} \lambda_i L_{\mathrm{i}} \tag{1}$$

$$L_{\mathrm{i}} = -\sum_{c=1}^{C_i} y_{ic} \log(\hat{\mathrm{y}}_{ic}) \tag{2}$$

In the formulas, N represents the total number of tasks, and λ_i is the weight coefficient for the i-th task. C_{i} is the number of categories in the i-th task. y_{ic} is the true label for category c in task i, and $\hat{\mathrm{y}}_{ic}$ is the corresponding predicted probability. This integrated loss function design ensures that the model optimizes multiple tasks simultaneously during the training process, thereby enhancing the overall performance and generalization capability of the model.

3.2. Multi-Label Learning with Joint Feature and Label Attention Mechanism

Image features and label embeddings are considered cross-modal features. We have designed a multi-label learning strategy based on joint feature and label attention to enhance the model's utilization of the combined features of image characteristics and label embeddings. This approach involves learning the cross-modal correlations between image feature information and label embeddings, deeply integrating multi-regional image

features into the label embeddings. Additionally, this process includes learning the semantic co-occurrence among labels. The structure of this mechanism is illustrated in Figure 2.

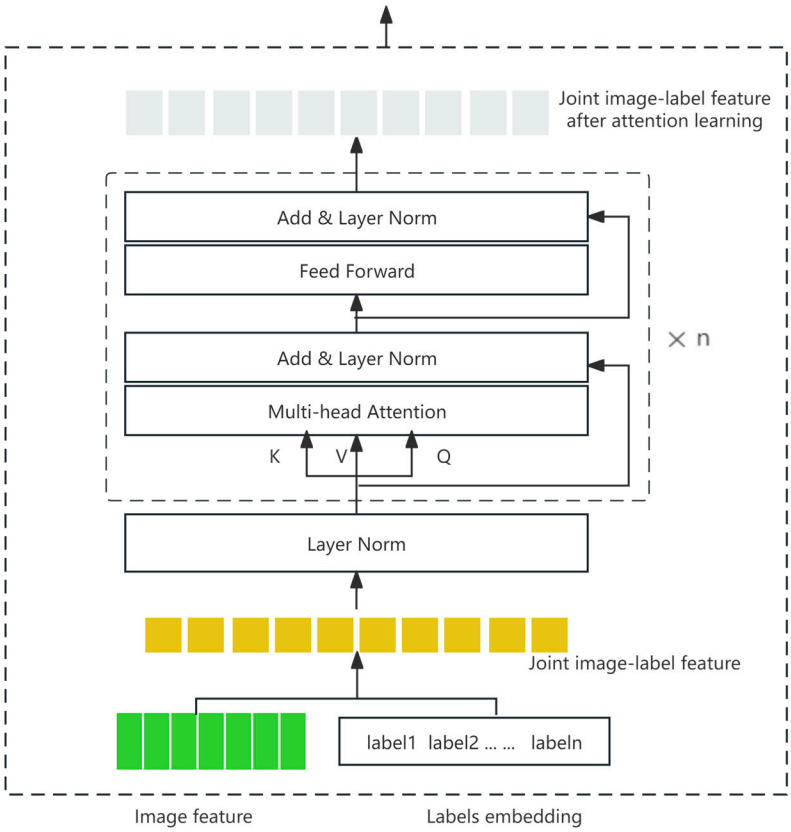

Figure 2. The structure of joint feature and label attention mechanism. Image features and label embeddings are concatenated to form a joint image label feature representation, which is then input into the joint feature attention module. Before inputting, the joint features should be normalized to enhance the stability of the data features. The subsequent attention module structure is identical to the encoder structure in the transformer [21] and can be stacked in multiple layers. After attention learning, the joint feature obtained includes associations between image features and label embeddings, as well as co-occurrences of feature semantics among various label embeddings. This feature is used for multi-label classification prediction by a subsequent classifier.

The joint representation of image labels is denoted by F, $F \in R^{(H' \times W' + N) \times C'}$, where H' represents the height of the feature map, W' is the width, C' is the number of channels, and N is the number of multi-label categories. After undergoing attention learning, $F' \in R^{(H' \times W' + N) \times C'}$, the joint feature with cross-modal attention information can be represented by the following formula:

$$F' = Concat(head_1, head_2, \ldots\ldots head_{\mathrm{h}})W^O \tag{3}$$

$$\mathrm{head}_i = softmax(\frac{FW_i^Q(FW_i^K)^T}{\sqrt{\mathrm{d}_k}})FW_i^V \tag{4}$$

In this formula, h represents the number of attention heads, W^O is the learnable weight parameter within the multi-head attention, $\sqrt{\mathrm{d}_k}$ serves as the scaling factor, and W_i^Q, W_i^K, and W_i^V are the learnable weight parameters for the given inputs.

To mitigate the impact of imbalanced label distribution in samples, this study employs an Asymmetric Loss function (ASL) [22] to enhance the performance of multi-label classification under conditions of label imbalance.

$$ASL = \frac{1}{K}\sum_{k=1}^{K}\begin{cases}(1-p_k)^{\gamma^+}\log(p_k), y_k = 1\\ (p_{m,k})^{\gamma^-}\log(1-p_{m,k}), y_{k=0}\end{cases} \quad (5)$$

$$p_{m,k} = \max(p_{k-m,0}) \quad (6)$$

In this context, y_k represents the true label, p_k is the predicted value, K is the total number of categories, and γ^+ and γ^- are the modulation indices for the contribution of positive and negative sample losses to the total loss, respectively. $p_{m,k}$ is a probability shift used to mitigate the contribution of simple negative samples to the overall loss, and m is the modulation parameter for this probability shift.

The entire image processing module encompasses multitask-based pretraining and multi-label learning based on feature fusion and label attention. The architecture of this module is illustrated in Figure 3:

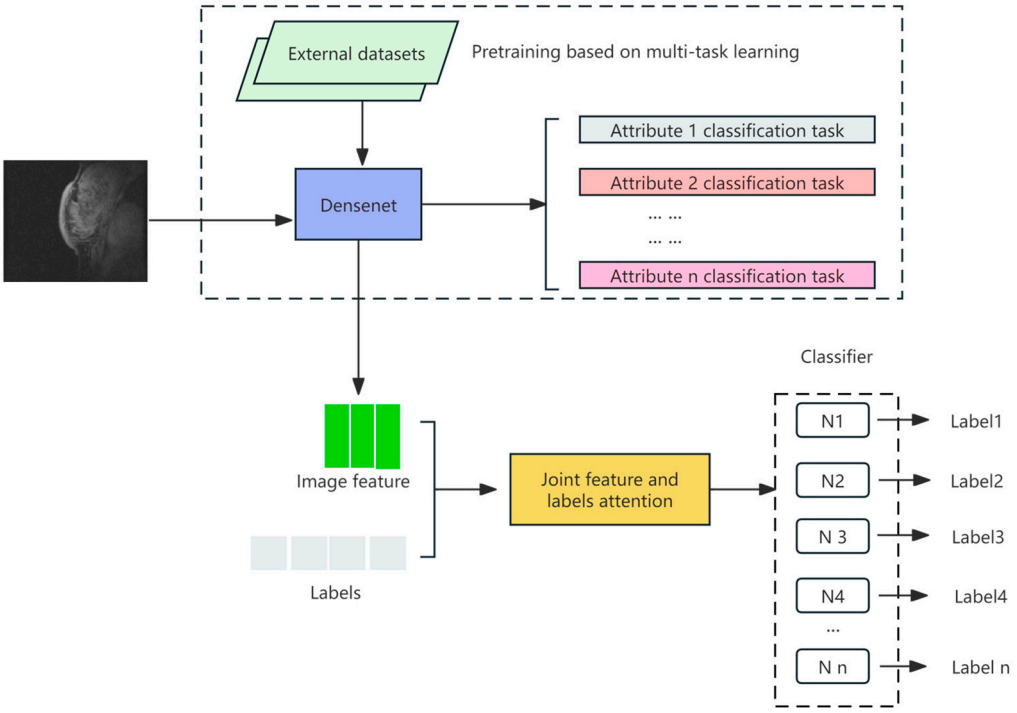

Figure 3. The architecture of multitask-based pretraining and multi-label learning.

3.3. Prompting LLMs to Answer Question

In Med-VQA, introducing LLMs to process and answer questions is a key step to enhance the overall system performance. Leveraging the powerful language comprehension and generation capabilities of LLMs enables an effective understanding of the intent and direction of questions. This section explains how to activate LLMs for Med-VQA tasks using prompting techniques, and how to integrate this technology into the ITLTA framework to optimize the performance of Med-VQA.

In Med-VQA, the format of questions is relatively straightforward, making it easy for a frozen large language model (LLM) to understand their intent. For instance, typical questions in the data might include, "What part of the body is being imaged?", "Is this a CT scan?", or "What is abnormal in the MRI?" with concise answers such as "Breast", "Yes", and "Quadrilateral space syndrome". Therefore, we believe that at this stage in Med-VQA tasks, existing LLMs are sufficient for zero-shot learning to address these questions without the need for additional training in natural language understanding. We combine the visual

labels obtained from multi-label learning with simple prompts and a few example question–answer pairs, inputting them into the LLM. Experiments have demonstrated that the LLM can effectively produce the required answers.

3.4. System Interpretability

Since our framework initially obtains visual labels and then inputs these labels along with the medical image and question into the LLM, leveraging its natural language understanding capabilities to derive answers, we can verify the accuracy of the visual labels in the first phase during testing. If an error occurs, such as the model incorrectly identifying a visual label—for instance, labeling an image of the brain as a limb—the large language model is likely to provide an incorrect answer. Such mistakes allow us to easily pinpoint and explain errors in image understanding. If it happens, it demonstrates exactly where the system's understanding was incorrect. Conversely, if all visual labels or those relevant to the question are correct but the final answer is still incorrect, it can be easily explained as a misunderstanding of the question by the LLM. The clear identification of error sources significantly enhances the interpretability of the system.

4. Experiments and Results

In this section, we evaluate our system on the VQA-Med 2019 dataset. Additionally, we conducted ablation studies to verify the effects of pretraining on medical images and the performance of various LLMs.

4.1. Datasets

We compiled a pretraining dataset of approximately 140,000 medical images based on five publicly available medical image datasets. These datasets are ABIDE [23], ChestX-ray8 [24], LIDC-IDRI [25], MURA [26], and MIAS [27]. These datasets include modalities such as CT, MRI, X-ray, and ultrasound, and cover various organ categories including the brain, lungs, breasts, limbs, and shoulders. They also include corresponding lesion labels. We categorized the data based on three attributes (modality, organ, and presence of lesions) to facilitate subsequent multitask pretraining.

The VQA-Med 2019 dataset [6] was introduced during the ImageCLEF 2019 challenge. This dataset primarily encompasses four major categories: modality, plane, organ system, and abnormality. It includes a training set consisting of 3200 medical images with 12,792 question–answer (QA) pairs, a validation set comprising 500 medical images with 2000 QA pairs, and a test set featuring 500 medical images with 500 QA pairs.

We categorized each image in the training data based on the answers to the questions, turning every image into a multi-labeled medical image. For many "yes or no" questions, we also organized corresponding attribute labels based on the answers. The attributes of modality, organ, and plane had fewer subclass labels, while there were too many different types of abnormality labels. However, since the training data themselves are limited, many categories of abnormalities only have 1–2 instances, which has been a reason for lower accuracy in previous studies. We organized the entire dataset into a multi-labeled dataset based on the answers to the questions, preparing it for multi-label training.

4.2. Implementation Details and Training

We selected Densenet201 as the backbone network, which was pretrained on the ImageNet dataset. We then proceeded with multitask pretraining for medical images. Our pretraining dataset comprised a total of 143,353 medical images. The training was organized around three main attributes: modality, organ system, and presence of abnormalities. The training and validation sets were split in an 80% to 20% ratio. Images were resized to 224×224 before training. We used the PyTorch deep learning framework and trained the model on four NVIDIA Tesla V100 GPUs for 100 epochs.

Subsequently, we conducted multi-label learning training based on feature and label joint attention mechanisms using the pretrained model. We set the values for γ^+ and γ^- to

0 and 4, respectively, to enhance the contribution of negative samples to the overall loss. The probability shift parameter m was set to 0.05. We configured the model with 4 attention heads, a batch size of 256, and ran the training for 100 iterations.

During the question-answering phase with LLMs, we input four components into the model: (1) the predicted set of visual labels, (2) the question, (3) our prompt text, and (4) ten example question–answer pairs to serve as references for the LLM. Experiments have demonstrated that these components are sufficient for the LLMs to understand and respond to the questions accurately. We utilized two open-source large language models, GLM-6B [28] and Baichuan-13B [29], for our experiments.

4.3. Ablation Studies

To assess the effectiveness of each component within our framework, we conducted detailed ablation studies. These experiments were designed to evaluate the impact of various configurations on performance: (1) multi-label learning without the use of external medical image datasets for pretraining; (2) multi-label learning conducted post-pretraining with external medical image datasets; (3) testing the efficacy of different large language models; specifically, we evaluated two prominent models, GLM-6B and Baichuan-13B.

As the test set was also categorized into the four major groups of modality, organ, plane, and abnormality, we initially conducted tests on the multi-label learning results using this dataset. To simplify the analysis, we calculated the accuracy for each category based on a single-label approach to evaluate the model's performance in each category. The results of the ablation study for multi-label learning, with and without pretraining, are shown in Table 1. Accuracy(O), Precision(O), Recall(O), and F1(O) represent the results without pretraining, and Accuracy(W), Precision(W), Recall(W), and F1(W) represent the results with pretraining. The table illustrates that pretraining with a large volume of medical images significantly enhances image recognition accuracy, particularly evident in the less represented category of abnormality.

Table 1. The results of multi-label learning with/without external medical pretraining.

	Modality	**Plane**	**Organ**	**Abnormality**
Accuracy(O)	0.784	0.808	0.736	0.136
Precision(O)	0.470	0.514	0.592	0.095
Recall(O)	0.677	0.554	0.550	0.134
F1(O)	0.555	0.533	0.571	0.111
Accuracy(W)	0.856	0.864	0.840	0.216
Precision(W)	0.518	0.520	0.655	0.102
Recall(W)	0.786	0.613	0.698	0.224
F1(W)	0.625	0.562	0.676	0.140

We also conducted ablation studies using different large language models, compiling the results with and without pretraining into Table 2. GLM(O) and Baichuan(O) represent the outcomes of these two large language models without the use of pretraining, while GLM(W) and Baichuan(W) show their results with pretraining. The ablation study results across different models are very similar. Analyzing the responses, we found that most errors were due to incorrect image visual labels leading to incorrect answers. Conversely, errors solely caused by the large language models were minimal, and the results between the two models showed little variance. This suggests that the current language understanding capabilities of large language models are sufficient for Medical Visual Question-Answering tasks. However, the significant differences in results with and without pretraining align with trends observed in multi-label learning, indicating that these discrepancies stem from the impact of medical image pretraining on feature extraction.

Table 2. The results with/without external medical pretraining of different LLMs.

LLMs	Modality	Plane	Organ	Abnormality	Overall
GLM(O)	0.768	0.800	0.712	0.184	0.616
Baichuan(O)	0.744	0.792	0.704	0.176	0.604
GLM(W)	0.848	0.840	0.824	0.280	0.698
Baichuan(W)	0.840	0.832	0.808	0.280	0.690

4.4. Comparison with State of the Art

Following our ablation studies, we chose the best-performing model from our experiments, represented as ITLTA (GLM), to compare against various state-of-the-art methods on the VQA-Med 2019 dataset. The results, as shown in Table 3, indicate that our comprehensive results surpass the best-performing model by 0.8%. Additionally, our method performed well across multiple categories. However, our results did not excel in the "plane" category compared to MMBERT(P) and WSDAN, primarily because our pretraining data did not include data specific to the "plane" category, leading to a less prominent performance in this aspect. Overall, our approach, which leverages enhanced image understanding capabilities and the linguistic capabilities of large language models, has increased the feasibility of Med-VQA applications in clinical settings.

Table 3. Comparison between the experimental results of this study and the results of existing methods.

Methods	Modality	Plane	Organ	Abnormality	Overall
TUA1 [30]	0.667	0.716	0.744	0.035	0.606
CGMVQA [31]	0.805	0.808	0.728	0.017	0.600
MMBERT(NP)	0.806	0.816	0.712	0.043	0.602
MMBERT(P)	0.833	0.864	0.768	0.140	0.672
WSDAN(P)	0.847	0.864	0.768	0.201	0.690
ITLTA(GLM)	0.848	0.840	0.824	0.280	0.698

5. Conclusions

The Image to Label to Answer (ITLTA) framework proposed in this paper offers a novel approach to Med-VQA. By effectively merging multi-label learning of medical images with the deep linguistic insights provided by LLMs, this framework substantially reduces the dependence on extensive training datasets traditionally necessary for Med-VQA. Additionally, the ITLTA framework enhances the interpretability of the Med-VQA process through the intermediate visual representation of labels and prompts, fostering transparency that is crucial for increasing clinicians' trust and acceptance of the model predictions. Future developments for the ITLTA framework could include the extension of the diversity of external medical image datasets, which would improve performance across a wider range of categories without the need for specific dataset customizations for Med-VQA. Such adaptability makes the clinical application of the ITLTA framework more flexible, making it suitable for a broader array of medical question-answering scenarios and demands.

Author Contributions: Conceptualization, K.P.S., J.W. and L.-M.A.; methodology, J.W., Y.S. and K.P.S.; resources, K.P.S.; data curation, J.W., K.P.S., Y.S., D.H. and L.-M.A.; writing—original draft preparation, K.P.S., J.W. and L.-M.A.; writing—review and editing, K.P.S., J.W. and L.-M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in this article. This study utilized the publicly available Medical Visual Question Answering dataset VQA-Med 2019. This dataset was openly collected and released under appropriate ethical standards and frameworks for data protection, specifically designed for scientific research purposes. In this study, we ensured

that all data handling and analysis activities strictly adhered to the usage conditions and privacy protection guidelines set by the dataset providers.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
2. Nguyen, B.D.; Do, T.-T.; Nguyen, B.X.; Do, T.; Tjiputra, E.; Tran, Q.D. Overcoming data limitation in medical visual question answering. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part IV 22. pp. 522–530.
3. Gupta, D.; Suman, S.; Ekbal, A. Hierarchical deep multi-modal network for medical visual question answering. *Expert Syst. Appl.* **2021**, *164*, 113993. [CrossRef]
4. Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; Ge, Z. Medical visual question answering: A survey. *Artif. Intell. Med.* **2023**, *143*, 102611. [CrossRef] [PubMed]
5. Al-Sadi, A.; Al-Ayyoub, M.; Jararweh, Y.; Costen, F. Visual question answering in the medical domain based on deep learning approaches: A comprehensive study. *Pattern Recognit. Lett.* **2021**, *150*, 57–75. [CrossRef]
6. Abacha, A.B.; Hasan, S.A.; Datla, V.V.; Liu, J.; Demner-Fushman, D.; Müller, H. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In Proceedings of the CLEF (Working Notes), Lugano, Switzerland, 9–12 September 2019; Volume 2.
7. Hasan, S.A.; Ling, Y.; Farri, O.; Liu, J.; Müller, H.; Lungren, M.P. Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In Proceedings of the CLEF (Working Notes), Avignon, France, 10–14 September 2018.
8. Liu, F.; Peng, Y.; Rosen, M.P. An effective deep transfer learning and information fusion framework for medical visual question answering. In Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland,, 9–12 September 2019; Proceedings 10. pp. 238–247.
9. Ambati, R.; Dudyala, C.R. A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering. In Proceedings of the 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 16–18 December 2018; pp. 1–6.
10. Zhan, L.-M.; Liu, B.; Fan, L.; Chen, J.; Wu, X.-M. Medical visual question answering via conditional reasoning. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2345–2354.
11. Gong, H.; Chen, G.; Liu, S.; Yu, Y.; Li, G. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 456–460.
12. Pan, H.; He, S.; Zhang, K.; Qu, B.; Chen, C.; Shi, K. Muvam: A multi-view attention-based model for medical visual question answering. *arXiv* **2021**, arXiv:2107.03216.
13. Huang, X.; Gong, H. A Dual-Attention Learning Network with Word and Sentence Embedding for Medical Visual Question Answering. *arXiv* **2023**, arXiv:2210.00220. [CrossRef] [PubMed]
14. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
15. Moon, J.H.; Lee, H.; Shin, W.; Kim, Y.-H.; Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 6070–6080. [CrossRef] [PubMed]
16. Hartsock, I.; Rasool, G. Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review. *arXiv* **2024**, arXiv:2403.02469.
17. Khare, Y.; Bagal, V.; Mathew, M.; Devi, A.; Priyakumar, U.D.; Jawahar, C. Mmbert: Multimodal bert pretraining for improved medical vqa. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 1033–1036.
18. Kalyan, K.S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2023**, *6*, 100048. [CrossRef]
19. Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N.A.; Luo, J. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 2963–2975.
20. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
22. Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric loss for multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 82–91.

23. Craddock, C.; Benhajali, Y.; Chu, C.; Chouinard, F.; Evans, A.; Jakab, A.; Khundrakpam, B.S.; Lewis, J.D.; Li, Q.; Milham, M. The neuro bureau preprocessing initiative: Open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* **2013**, *7*, 5.
24. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
25. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef] [PubMed]
26. Rajpurkar, P.; Irvin, J.; Bagul, A.; Ding, D.; Duan, T.; Mehta, H.; Yang, B.; Zhu, K.; Laird, D.; Ball, R.L. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv* **2017**, arXiv:1712.06957.
27. Suckling, J.; Parker, J.; Dance, D.; Astley, S.; Hutt, I.; Boggis, C.; Ricketts, I.; Stamatakis, E.; Cerneaz, N.; Kok, S. *Mammographic Image Analysis Society (Mias) Database v1.21*; Apollo—University of Cambridge Repository: Cambridge, UK, 2015.
28. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. Glm: General language model pretraining with autoregressive blank infilling. *arXiv* **2021**, arXiv:2103.10360.
29. Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D. Baichuan 2: Open large-scale language models. *arXiv* **2023**, arXiv:2309.10305.
30. Zhou, Y.; Kang, X.; Ren, F. TUA1 at ImageCLEF 2019 VQA-Med: A Classification and Generation Model based on Transfer Learning. In Proceedings of the CLEF (Working Notes), Lugano, Switzerland, 9–12 September 2019.
31. Ren, F.; Zhou, Y. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access* **2020**, *8*, 50626–50636. [CrossRef]