MLeVLM: Improve Multi-level Progressive Capabilities based on Multimodal Large Language Model for Medical Visual Question Answering

Dexuan Xu[1†], Yanyuan Chen[2†], Jieyi Wang[2], Yue Huang[2], Hanpin Wang[1], Zhi Jin[1], Hongxing Wang[4], Weihua Yue[5], Jing He[6], Hang Li[7], Yu Huang[3*]

[1] School of Computer Science, Peking University
[2] School of Software & Microelectronics, Peking University
[3] National Engineering Research Center for Software Engineering, Peking University
[4] Xuanwu Hospital Capital Medical University, [5] Peking University Sixth Hospital
[6] Peking University People's Hospital, [7] Peking University First Hospital

Abstract

Medical visual question answering (MVQA) requires in-depth understanding of medical images and questions to provide reliable answers. We summarize multi-level progressive capabilities that models need to focus on in MVQA: **recognition, details, diagnosis, knowledge, and reasoning**. Existing MVQA models tend to ignore the above capabilities due to unspecific data and plain architecture. To address these issues, this paper proposes **M**ulti-**le**vel Visual Language Model (MLeVLM [1]) for MVQA. On the data side, we construct a high-quality multi-level instruction dataset MLe-VQA via GPT-4, which covers multi-level questions and answers as well as reasoning processes from visual clues to semantic cognition. On the architecture side, we propose a multi-level feature alignment module, including attention-based token selector and context merger, which can efficiently align features at different levels from visual to semantic. To better evaluate the model's capabilities, we manually construct a multi-level MVQA evaluation benchmark named MLe-Bench. Extensive experiments demonstrate the effectiveness of our constructed multi-level instruction dataset and the multi-level feature alignment module. It also proves that MLeVLM outperforms existing medical multimodal large language models.

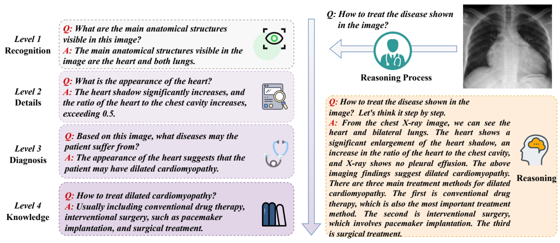


Figure 1: An illustration of multi-level reasoning process in medical visual question answering.

1 Introduction

Medical visual question answering (MVQA) is an interdisciplinary problem that combines CV and NLP which requires deep understanding of medical images and questions to provide reliable answers. Recently, lots of works have constructed professional MVQA datasets (He et al., 2020; Lau et al., 2018; Ben Abacha et al., 2021; Liu et al., 2021). However, limited by the number of parameters, the past VQA models usually focus on a specific dataset and cannot cover as many scenarios as possible. With the rapid development of large language models, multimodal large language models (MLLMs) have become an important solution for solving VQA tasks (Liu et al., 2023; Alayrac et al., 2022; Driess et al., 2023), due to the abundant number of parameters, large amount of pre-trained contents and excellent text generation capabilities. Similarly, by directly substituting the training data, these general domain methods have been directly transferred to the medical domain (Li et al., 2023a; Moor et al., 2023; Tu et al., 2023a).

More challenging than the general domain VQA is that when doctors conduct MVQA in real scenarios (Wu et al., 2022), they first need to focus on the key parts in the image, and further consider the detailed information afterward. Based on the visual clues acquired, they make a diagnosis of the disease, and then further answer knowledge-based questions such as how to treat it. We summarize these progressive capabilities in the following five aspects: **Recognition, Details, Diagnosis, Knowledge and Reasoning**. They cover a cascade of capabilities from low-level vision to high-level semantics, simulating the thought process of a doctor in answering a medical question. As is illustrated in Figure 1, for a complex semantic question (*"How to treat the disease shown in the image?"*), obtaining its answer is not straightforward, but involv-

* Corresponding author: Yu Huang, hy@pku.edu.cn.
†These authors contributed equally to this work.

[1]MLeVLM is going to be publicly available at https://github.com/RyannChenOO/MLeVLM

Findings of the Association for Computational Linguistics ACL 2024, pages 4977–4997
August 11-16, 2024 ©2024 Association for Computational Linguistics

ing a sequence of implicit complex reasoning or medical knowledge. That is, recognize the objects (*"What are the main anatomical structures visible in this image?"*), extract local details (*"What is the appearance of the heart?"*), diagnose based on acquired visual clues (*"Based on this image, what diseases may the patient suffer from?"*), consider knowledge of specific diseases (*" How to treat dilated cardiomyopathy?"*), and finally integrate the information obtained from the above-mentioned processes to form a reasoning process.

However, existing MVQA models do not have the multi-level progressive capabilities and suffer from the following two problems: 1) Unspecific data. The public training data (Moor et al., 2023) used by the model is often derived from paper crawlers and lacks instruction data for specific construction, especially for multi-level progressive capabilities. 2) Plain architecture. Existing multimodal large language models usually align visual features and text features through simple linear layers (Li et al., 2023a). Since different capabilities focus on different modal features, it is difficult to effectively align multimodal feature with a simple linear layer. Therefore, in this paper, we propose **M**ulti-**le**vel Visual Language Model for MVQA (MLeVLM), which aims to improve the multi-level capabilities in medical visual question answering. On the data side, we first carefully construct a dataset containing 60K high-quality multi-level instructions called **MLe-VQA**. The dataset covers multi-level progressive capabilities: recognition, details, diagnosis, knowledge and reasoning. On the architecture side, we propose a multi-level feature alignment module to effectively align features from visual to semantic. Specifically, it contains an attention-based token selector that focuses on important areas and a context merger that fuses visual details. To better evaluate the model's capabilities of different levels, we manually construct a multi-level evaluation benchmark named **MLe-Bench**. Experimental results on the benchmark demonstrate the effectiveness of our constructed multi-level instruction dataset and the multi-level feature alignment module. It also proves that MLeVLM outperforms existing medical MLLMs.

Our contributions are as follows:

1) We use GPT-4 to carefully construct the multi-level instruction dataset MLe-VQA. It ensures that the model can fully get the capabilities of recognition, details, diagnosis, knowledge and reasoning.

2) We design a multi-level feature alignment module for MLeVLM including attention-based token selector and context merger to efficiently align different levels of features from visual to semantic.

3) We manually construct a multi-level evaluation benchmark called MLe-Bench from existing public VQA datasets. The results on the benchmark demonstrate the effectiveness of our proposed MLeVLM from both data and architecture sides.

2 Related Work

2.1 Multimodal Large Language Models

Motivated by LLM's outstanding capabilities, researchers are exploring ways to transfer these capabilities to the vision domain, developing multimodal LLM (Alayrac et al., 2022; Liu et al., 2023; Chen et al., 2023a; Bai et al., 2023; Ye et al., 2023). Existing research on MLLMs aims to link visual encoders with large language models, where the former is responsible for visual perception and the latter for semantic understanding. Flamingo (Alayrac et al., 2022) is one of the earlier works. It uses a vision encoder to extract visual embeddings, a vision-language resampler module to connect the vision encoder to a frozen language model, and employs multi-layer cross-attention to fuse multimodal inputs. BLIP2 (Li et al., 2023c) uses a Q-Former to connect the frozen LLM and vision encoder. MiniGPT-4 (Chen et al., 2023a) freezes the parameters of the vision encoder and LLM, and only optimizes a trainable projection matrix to connect the vision and language layers. LLaVA (Liu et al., 2023) simply uses a linear projection layer and it freezes the vision encoder then training LLM during the instruction tuning phase. In contrast, mPLUG-owl (Ye et al., 2023) trains the vision encoder in the pre-training stage while freezing the LLM to align visual and language embeddings, and subsequently, in the instruction tuning stage, it freezes the vision encoder and trains the LLM. Appendix A.1 and Appendix A.2 show more related work about construction of the MLLM instruction dataset and reasoning in MLLM, respectively.

2.2 Medical MLLM

The advancements in generic MLLM have also contributed to the process of MLLM in the biomedical field (Zhou et al., 2023b; Guo et al., 2023; Zhou et al., 2023a). A recent work, LLaVA-Med (Li et al., 2023b) extracts biomedical image-text pairs from PubMed Central and uses GPT-4 (OpenAI, 2023) to self-instruct biomedical multimodal instruction-following data. Utilizing a substantial amount of image-text pairs, LLaVA is fine-tuned to align image-text tokens to the biomedical domain. Subsequently, the model is trained on the aforementioned custom dataset, enabling it to perform biomedical instructions and various tasks in a conversational manner. Med-PaLM (Tu et al., 2023b) fine-tunes and aligns PaLM-E (Driess et al., 2023) to the biomedical domain, using a multi-task, multi-modal medical dataset with more than 1 million samples for training and evaluation. Based on the open source OpenFlamingo (Awadalla et al., 2023) and constructed pre-trained medical data, Med-Flamingo (Moor et al., 2023) migrates in-context learning and few-shot learning abilities of Flamingo to the medical domain. By employing few-shot prompting, users are allowed to customize the response format, e.g., to provide rationales for the given answers, but few-shot multimodal prompted rationales may not be robust. We list the relationship between our method and previous work in Appendix A.3.

3 Construction of Multi-level Dataset and Benchmark

3.1 Multi-level Capabilities in Medical VQA

Medical visual question answering is a complex task that requires the understanding of visual clues and semantic cognition. In tackling such complex tasks, especially when answering advanced semantic questions, involves an intrinsic requirement for sequential and progressive reasoning by integrating observed visual cues. In this paper, we outline the five progressive capabilities essential for MLLM in handling MVQA tasks, aiming to simulate the logical process of a human doctor when responding to medical visual questions.

1) **Recognition.** Basic visual perception involves overall observation of images for object recognition and initial comprehension of fundamental features, e.g., basic anatomical structures and imaging modalities.

2) **Details.** Further detailed observation requires the capability to capture subtle features and correlations among objects in the image, e.g., organ size, color, and tissue properties.

3) **Diagnosis.** Identifying abnormalities and diagnosing diseases from the image requires synthesizing the interpretation of visual clues obtained above and conducting in-depth analysis through incorporating external knowledge.

4) **Knowledge.** Focusing on higher-level knowledge-related questions regarding the diagnosis results, e.g., the treatment and prevention of diseases, requiring the integration of out-of-image knowledge to provide recommendations.

5) **Reasoning.** Reasoning builds on the capabilities of the previous levels. The above four capabilities exhibit a hierarchical relationship and this progressive question grading structure can be used to guide MLLM in comprehending the image step by step, explicitly providing the necessary information for the reasoning process.

3.2 MLe-VQA Dataset

Currently there is a lack of multimodal medical datasets to train an assistant with the aforementioned capabilities. To fill this gap, we construct the first comprehensive **M**ulti-**le**vel medical VQA dataset, called MLe-VQA. It consists of medical images spanning multiple modalities from public datasets, along with high quality questions and answers constructed using the GPT-4 and self-instruction methods. Specifically, the dataset contains 6K images and 60K Q&A pairs. Figure 2 shows an outline of its construction process.

Data Source. To maximise the diversity and comprehensiveness of our data, we compile about 6K images from multiple public data sources. More details can be found in the Appendix B.2. We collect public datasets that provide images and captions, using the captions of the images to prompt language-only GPT-4 for generating question-answer pairs.

Image Selection. Sometimes, image captions are too short for GPT-4 to generate meaningful questions and answers. To select clinical images with rich information, we initially screen out image-caption pairs where the caption is mainly describing the patient, and further discard some of the short-caption images based on the caption length.

GPT-Assisted Multi-level Data Generation. Given an image caption, we design prompt to make GPT-4 generate meaningful questions and answers in a tone as if it could see the image (even though it only has access to the text). As shown in Figure 2, within the instruction, we devised a base prompt for a general description to instruct the GPT-4 API about the specific task requirements. Additionally, we include a level-specific prompt to define each level. Building upon this foundation, the GPT-4

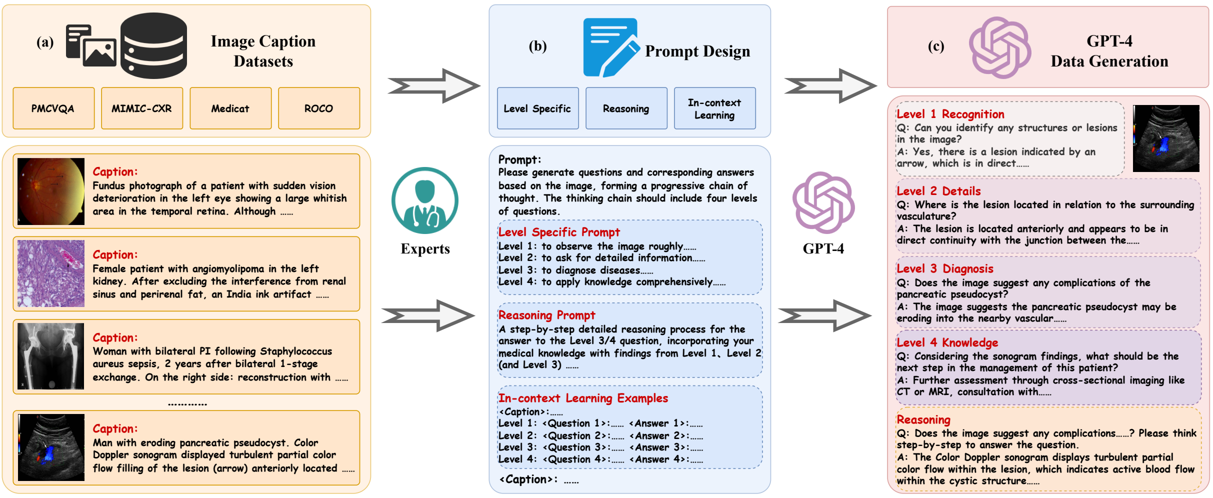


Figure 2: An illustration of the process to prompt GPT-4 to generate multi-level instructions.

is prompted to construct reasoning Q&A pairs by obtaining progressively complex image-related information. We also manually design several in-context learning examples for each level to ensure that the GPT-4 has a clearer understanding of the intent of each level and how to generate high-quality Q&A pairs based on the captions provided. Refer to the Appendix E.5 for more detailed examples of in-context learning. The final statistics of the dataset are shown in Appendix C.

3.3 MLe-Bench

Evaluating model capabilities at each level is highly difficult and there is a lack of evaluation benchmarks. Therefore, we construct MLe-Bench, a challenging multimodal problem set of total 1,492 questions to assess multi-level capabilities. Specifically, we invited experts with clinical knowledge to manually select the questions related to each level based on existing public medical VQA datasets. Table B.3 shows the statistics of the datasets, where the questions are extracted from the officially divided test set. To make the benchmark more practical for evaluating MLLM, we rephrase the answers to be open-ended. This makes the benchmark more challenging and realistic, as the models have to independently answer complex questions, rather than selecting the most reasonable answer choice from a limited set of choices.

4 Model Architecture

Existing multi-modal large language models usually align visual features and text features through simple linear layers. To effectively align features of different levels from visual to semantic in medical VQA, we design a multi-level feature alignment (MLFA) module. Specifically, we add attention-based token selector and context merger to the widely used visual encoder and text encoder to improve the model's focus on key areas and visual details. We then use two projectors to map the low-level visual information and high-level semantic information into the large language model. The model architecture is shown in Figure 3.

4.1 Visual Encoder and Text Encoder

The model first processes medical images and user interaction instructions through a visual encoder and a text encoder. For a given image input $I \in \mathbb{R}^{H \times W \times 3}$, we use the pre-trained vision transformer to get the visual embedding $X \in \mathbb{R}^{N \times C}$, where $N = HW/p^2$, C represents the number of embedded channels and p represents the patch size. For a given user instruction input T, we can generate text-guided query $Q_t \in \mathbb{R}^{K \times D}$ through the pre-trained text encoder and the resulting visual embedding X, where K denotes the number of queries and D denotes the dimension of queries. We choose Q-Former (Dai et al., 2023) as this cross-modal text encoder to ensure that Q_t contains more visual clues related to user instructions.

4.2 Attention-based Token Selector

For the visual recognition question, we hope that the model pays more attention to key areas in the image. Therefore, we design an attention-based token selector. Its goal is to estimate the importance of each image patch based on cross-modal attention weights and select the most informative part to represent the whole image. Specifically, for text query Q_t and visual embedding X, we can calcu-

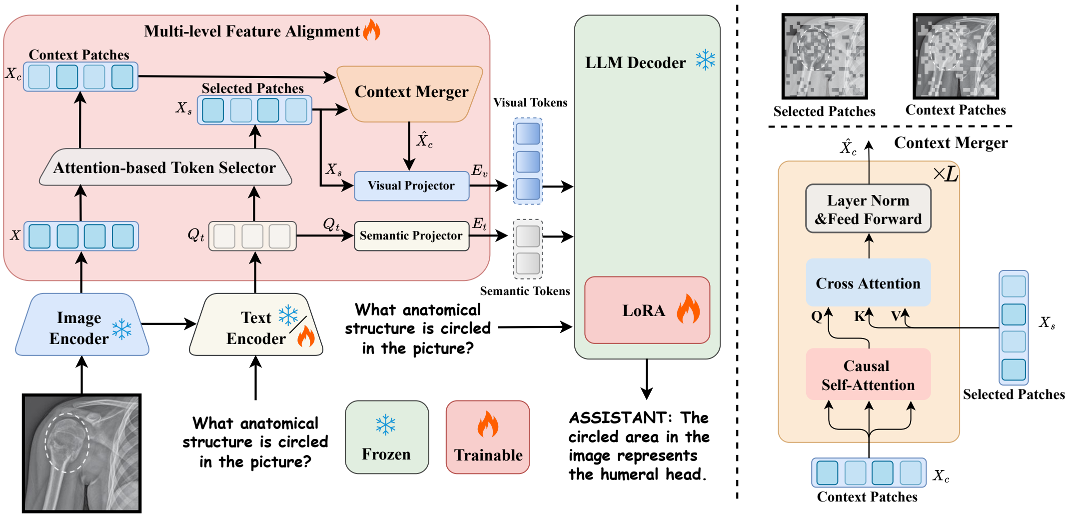


Figure 3: The model architecture of MLeVLM, which includes visual encoder, text encoder, the multi-level feature alignment module composed of the attention-based token selector and the context merger, and the LLM decoder. Details of the context merger are shown on the right.

late the attention weight matrix A through matrix multiplication:

$$A = \frac{\mathbf{W_q} Q_t \times (\mathbf{W_x} X^T)}{\sqrt{N}}, \tag{1}$$

where $\mathbf{W_q}$ and $\mathbf{W_x}$ are learnable parameters, $A \in \mathbb{R}^{K \times N}$. Inspired by (Rao et al., 2021; Jin et al., 2023), we predict cross-modal feature distribution $\pi \in \mathbb{R}^{N \times 2}$ with a lightweight module containing several MLP layers, where $\pi_i = \text{MLP}(A_i)$. By sampling from the distribution π, we can get a binary decision mask $M \in \{0,1\}^N$. To relax the sampling to be differentiable, we apply the Gumbel-Softmax trick (Maddison et al., 2016) to π:

$$\hat{\pi_{i,j}} = \frac{\exp\left(\left(\log \pi_{i,j} + g_{i,j}\right)/\tau\right)}{\sum_{r=1}^{2} \exp\left(\left(\log \pi_{i,r} + g_{i,r}\right)/\tau\right)}, \tag{2}$$

where $g_{i,j}$ are i.i.d samples drawn from Gumbel(0, 1) distribution[2], τ is the temperature parameter that controls the smoothness of softmax function. Afterwards, the binary decision mask M can be sampled from $\hat{\pi}$. Through this mask, we can get the selected key image patches $X_s \in \mathbb{R}^{P \times C}$ and the unselected context image patches $X_c \in \mathbb{R}^{(N-P) \times C}$, where P is the number of selected patches.

4.3 Context Merger

For the visual detail question, the contextual information of the image is as important as the key information in the image. Therefore, we fuse the contextual information and key information through the attention module instead of directly discarding the context. Specifically, our context fusion module consists of L stacked blocks, similar to the transformer decoder. Each block contains a causal self-attention layer, a cross-attention layer and a feed-forward layer[3]. The cross-attention layer uses X_c' as query, X_s as key and value to achieve context information fusion based on similarity in the feature space. The fused contextual feature $\bar{X}_c$ can be expressed as:

$$\bar{X}_c = softmax\left(\frac{\mathbf{W}_Q X_c' (\mathbf{W}_K X_s)^T}{\sqrt{C}}\right) \mathbf{W}_V X_s, \tag{3}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are the learnable parameter matrices, and X_c' is the result of the context information updated by the causal self-attention layer. After L-layer blocks, we can get the final contextual integration feature $\hat{X}_c$. By utilizing the token selector and context fusion module, we can dynamically get the key area features and contextual integration features of each image. Through concatenation, we can obtain the representation of complete visual features $X_v = [X_s, \hat{X}_c] \in \mathbb{R}^{N \times C}$.

4.4 Text Generation and Training Strategy

We map X_v and Q_t to the dimensions of the pretrained large language model through two projectors to obtain E_v and E_t, respectively:

$$E_v = \mathbf{W}_v X_v + b_v, E_t = \mathbf{W}_t Q_t + b_t, \tag{4}$$

[2] The Gumbel(0, 1) distribution can be sampled using inverse transform sampling by drawing $u \sim$ Uniform(0,1) and computing $g = -\log(-\log(u))$.

[3] Using causal self-attention helps convert 2D raster-ordered features in the image encoder into sequences with causal dependencies, improving consistency with textual tokens in LLM (Jin et al., 2023).

Table 1: The results of comparison with previous studies on four capabilities. Best and second-best results are shown in **bold** and underline, respectively. Rouge and Score refer to Rouge-L and GPT-score, respectively.

Models	Recognition		Details		Diagnosis		Knowledge		Average
	Rouge	Score	Rouge	Score	Rouge	Score	Rouge	Score	Score
Zero-shot with existing medical MLLM									
GPT4V (Achiam et al., 2023)	0.242	3.132	0.179	2.942	0.081	3.644	0.127	3.848	3.392
Med-flamingo (Moor et al., 2023)	0.293	1.893	0.286	2.371	0.277	2.278	0.243	2.257	2.200
LLaVA-Med (Li et al., 2023a)	0.335	2.744	0.348	2.621	0.303	3.039	0.296	2.996	2.918
Instruction-tuning with existing MLLM									
LLaMa-adapterV2 (Gao et al., 2023)	0.431	2.497	0.411	2.413	0.396	2.722	<u>0.373</u>	2.945	2.644
Minigpt-v2 (Chen et al., 2023a)	0.404	2.743	<u>0.436</u>	<u>2.764</u>	<u>0.402</u>	3.082	0.356	2.983	2.893
mPLUG-Owl (Ye et al., 2023)	0.349	2.613	0.342	2.658	0.321	3.105	0.293	3.059	2.859
Qwen-VL-Chat (Bai et al., 2023)	0.380	2.566	0.432	2.698	0.341	2.881	0.343	2.839	2.746
LLaMA-VID (Li et al., 2023d)	0.429	2.658	0.421	2.606	0.378	3.107	0.358	3.046	2.854
Our methods									
MLeVLM (MLFA)	<u>0.434</u>	<u>2.767</u>	<u>0.436</u>	**2.802**	0.392	<u>3.114</u>	0.364	<u>3.093</u>	<u>2.944</u>
MLeVLM (LoRA)	**0.449**	**2.807**	**0.437**	2.741	**0.404**	**3.118**	**0.389**	**3.114**	**2.945**

where $\mathbf{W}_v$, $\mathbf{W}_t$, b_v and b_t are learnable parameters, $E_v \in \mathbb{R}^{N \times D_{LLM}}$, $E_v \in \mathbb{R}^{K \times D_{LLM}}$, D_{LLM} is the input dimension of the pretrained large language model. After linear projection, both the visual information E_v and the semantic information E_t are converted into the language space. These visual and semantic tokens can be used in subsequent LLM to generate answers that respond to user instructions.

In order to stimulate the capabilities of LLM in medical visual question answering as much as possible, a reasonable training strategy is crucial. Considering the training cost, we divide the training process into three stages, namely the medical modality alignment, the medical instruction-tuning, and the level instruction-tuning.

Stage I: Medical modality alignment. There is a gap in the feature space of the pre-trained encoder and decoder. Therefore, we first need to pre-train the MLFA module on medical multimodal datasets to align the encoder and decoder. The data includes a large number of public medical image classification, medical image caption and medical VQA datasets, with a total of 1,700K samples. Detailed data sources are shown in Appendix B.1, and detailed instruction templates are shown in Appendix E.1 and Appendix E.2. During training, we keep the encoder and LLM weights frozen and only update the parameters of MLFA module. Through stage I, we can pre-train the encoder and decoder in medical modality alignment, allowing the model to initially gain understanding capabilities in the medical field.

Stage II: Medical instruction-tuning. To ensure that the model can follow instructions in diverse medical fields, it is necessary to fine-tune the model with medical instructions. Specifically, we use the instructions provided by LLaVa-Med to fine-tune the dataset. LLaVa-Med filters PMC-15M to obtain 60K samples, and generates multiple rounds of dialogue through GPT-4, containing rich medical instructions. During training, we keep the visual encoder frozen and update the parameters of the other components. After stage II, the model has preliminary medical VQA capabilities and can achieve good zero-shot results on public datasets.

Stage III: Level instruction-tuning. To improve the multi-level capabilities of medical MLLM, we add a level instruction-tuning stage. We use the MLe-VQA dataset to fine-tune the model, containing 60K samples. We implement two types of fine-tuning strategies. The first is to only update the parameters of MLFA layers and freeze the LLM. The second is to update the parameters of MLFA layers and use LoRA (Hu et al., 2021) to fine-tune the LLM parameters. After stage III, our model can effectively understand question characteristics of different levels with reasoning capability.

5 Experiments

5.1 Experimental Setup

Datasets. We use the MLe-Bench introduced in Section 3.3 for the zero-shot evaluation. The benchmark contains a test set of 1,492 samples compiled

Table 2: Ablation studies on MLe-Bench and the test set of MLe-VQA. Best results are shown in **bold**. We used Rouge-L to evaluate the results of each capability automatically.

Dataset	Training Strategy	Method	Seletor	Merger	Recognition	Details	Diagnosis	Knowledge	Average
MLe-Bench	Instruction-tuning	(a) w/o Selector	-	-	0.349	0.363	0.302	0.287	0.325
		(b) w/o Merger	✓	-	0.356	0.367	0.303	0.295	0.330
		(c) Ours	✓	✓	**0.372**	**0.374**	**0.315**	**0.313**	**0.344**
	Level-tuning	(d) w/o Selector	-	-	0.414	0.409	0.392	0.358	0.393
		(e) w/o Merger	✓	-	0.428	0.422	0.390	**0.367**	0.402
		(f) Ours	✓	✓	**0.434**	**0.436**	**0.392**	0.364	**0.407**
MLe-VQA	Instruction-tuning	(g) w/o Selector	-	-	0.385	0.341	0.287	0.256	0.317
		(h) w/o Merger	✓	-	0.388	0.348	0.290	**0.258**	0.321
		(i) Ours	✓	✓	**0.391**	**0.354**	**0.295**	**0.258**	**0.325**
	Level-tuning	(j) w/o Selector	-	-	0.434	0.409	0.352	0.319	0.379
		(k) w/o Merger	✓	-	0.442	0.413	0.358	0.320	0.383
		(l) Ours	✓	✓	**0.458**	**0.429**	**0.369**	**0.324**	**0.395**

from 4 public medical VQA datasets. We also perform held-in experiments on the test set of MLe-VQA to evaluate the effectiveness of each component. Details of both datasets are provided in the Appendices B.2 and B.3.

Merics. Previous works typically employ accuracy as evaluation metric, which scoring all available answers of the datasets to calculate classification accuracy. Since the ground truth of the benchmark has been rephrased, we follow (Li et al., 2023a) to employ Rouge-L (Lin, 2004) as automatic metrics. Additionally we follow (Zheng et al., 2023) to adopt BLEU (Papineni et al., 2002) and Sentence Similarity (Zhang et al., 2019) for automatic evaluation. To bridge the gap of traditional metrics, we follow LLaVA-Med (Li et al., 2023a) to use GPT-API for scoring. For different levels we set different evaluation prompts with a overall score on a scale of 0 to 5, where a higher score indicates better overall performance. Detailed prompts are provided in the Appendix E.6.

Implementation Details. We instantiate the model with the pre-trained EVA-G (Fang et al., 2023) for visual encoder and Q-Former (Dai et al., 2023) for text encoder. We choose Vicuna-7B (Chiang et al., 2023) as the default large language model. See Appendix D for more details.

5.2 Comparison with Existing Methods

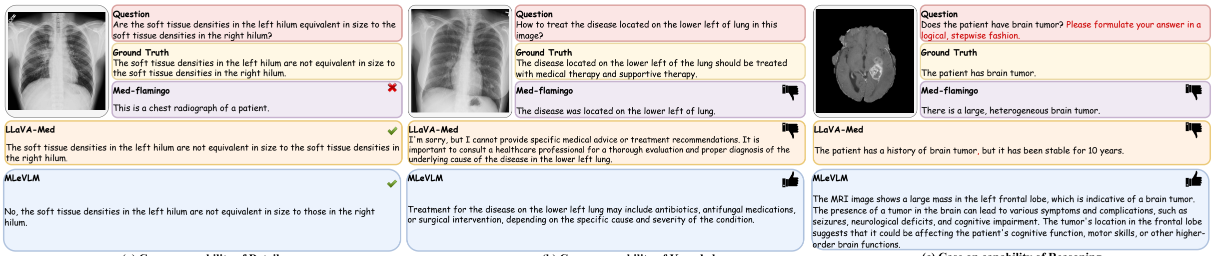
We perform comparative experiments of our method with existing medical MLLMs (Moor et al., 2023; Li et al., 2023a) and general MLLMs (Gao et al., 2023; Chen et al., 2023a; Ye et al., 2023; Bai et al., 2023; Li et al., 2023d). For existing medical MLLMs, we perform zero-shot testing on MLe-Bench directly. For general MLLM, we reproduce their models according to our full three-stage training strategy in Section 4.4 for a fair comparison. We use Rouge-L and GPT-score to evaluate the results of each capability. The experimental results are shown in Table 1. More results for traditional metrics are provided in the Appendix F.1.

From Table 1, we can observe that our methods achieve best evaluation results in all capabilities. This proves that MLeVLM is currently the best MVQA model in terms of four capabilities: Recognition, Details, Diagnosis and Knowledge. The model fine-tuned by Lora gets the best Rouge results on four levels and the best average GPT-score. Compared to the state-of-the-art open-source medical MLLM, our methods outperform in both text similarity and semantic correctness. Our methods also have better results compared to the MLLMs we reproduced. Since these methods are completely fine-tuned using the same data and only the model architecture is different, it suggest that our proposed model architecture is superior.

We also show the experimental results of using GPT4V to test directly on MLe-Bench. Based on the experimental results in the table below, it can be seen that although the API scoring of GPT4V will be higher than the existing MLLM model, none of the capabilities scored more than 4 points. In addition, the mean score for GPT4V was only 3.392. This proves that MLe-Bench is challenging and suitable for the existing 7B scale MLLM model for medical VQA evaluation.

5.3 Ablation Studies

We conduct ablation experiments on MLe-bench and the test set of MLe-VQA. Results are shown in Table 2. The contrasting model **w/o Selector**



(a) Case on capability of Details **(b) Case on capability of Knowledge** **(c) Case on capability of Reasoning**

Figure 4: Case studies on questions of different levels and reasoning capability.

discards token selector in the full model, and the contrasting model **w/o Merger** discards the context merger in the full model[4]. We reproduce the two models according to the training strategy in Section 4.4, and only fine-tune the MLFA layer during level instruction-tuning.

Ablation of Attention-based Token Selector. Comparing methods (d) and (e), it can be found that the metric of Recognition improves from 0.414 to 0.428. This indicates that token selector works in recognising the key areas. Similar conclusions can be found on instruction-tuning stage and the test set of MLe-VQA. This proves that the selector works in both stages.

Ablation of Context Merger. Comparing methods (e) and (f), it can be found that the metric of Details improves from 0.422 to 0.436. This indicates that integrating contextual information through context merger can assist models in recognizing visual details. Comparable findings have been observed in both the instruction-tuning stage and the test set of MLe-VQA, demonstrating the effectiveness of the merger across different stages.

Ablation of Level-tuning. Comparing methods (c) and (f), the model's capabilities at all four levels are significantly improved by adding level-tuning, indicating the validity of our constructed MLe-VQA.

Methods (c) and (f) achieve the relative best metrics, both in the instruction-tuning stage and in the level-tuning stage. This proves that our complete model can be adequately adapted to multi-level MVQA. Similar conclusions can be found on the test set of MLe-VQA. This proves that our model can be adapted to data with different distributions.

5.4 Analysis of Reasoning Capability

Since reasoning involves a long chain of thought process, we employ GPT and human evaluation to systematically evaluate the performance in reasoning. The results of GPT-score are presented in Figure 5, including the scores of reasoning on Diagnosis and Knowledge at MLe-Bench. The comparison methods include Med-flamingo, LLaVA-Med, and MLeVLM without level-tuning. It can be observed that: 1) By adding reasoning prompts, scores increased for LLaVA-Med and MLeVLM while decreased for Med-flamingo and MLeVLM w/o level-tuning. This suggests that LLaVA-Med and MLeVLM have reasoning capability and MLeVLM's reasoning capability comes from level-tuning; 2) On both levels of reasoning, MLeVLM's average scores improved by 0.061 and 0.025, respectively, which is higher than LLaVA-Med, demonstrating better reasoning capability. More analysis can be found in Appendix F.2.

In the course of **human evaluation**, annotators are required to grade each rationale on the criteria of Relevance, Correctness, Completeness, Coherence, and Explainability. Noted that although LLaVA-Med achieve decent results in terms of Relevant and Correct, its performance is poor than MLeVLM (both with and without level-tuning) in other aspects. In addition, after level tuning, our

Table 3: The results of human evaluation on MLe-Bench.

Method	Relevant	Correct	Complete	Coherent	Explainable
Med-flamingo	2.26	1.13	1.32	1.22	1.35
LLaVA-Med	4.18	2.50	3.69	3.80	3.63
Ours(Instruction-tuning)	3.84	2.50	4.20	4.30	4.20
Ours(Level-tuning)	**4.29**	**2.72**	**4.57**	**4.50**	**4.38**

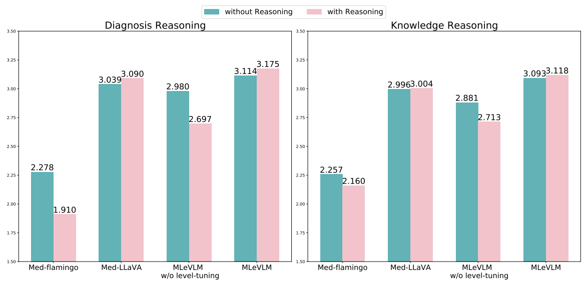


Figure 5: Comparison with existing medical MLLMs in reasoning capability. Use GPT-score for evaluation.

[4]Since the context merger is highly bound to the token selector, while discarding the token selector, the context merger is discarded together.

method significantly surpass other methods across all aspects of human evaluation, which is more valuable for interpretable studies in the medical field.

6 Qualitative Analysis

Case Study on Different Levels. To compare the capabilities between MLeVLM and existing medical MLLMs, we conduct case study on questions of different levels. Plot (a) of Figure 4 shows the models' answers to the detail-based questions. It can be seen that both LLaVA-Med and MLeVLM generate correct answers, while Med-flamingo has errors in its answers. Plot (b) of Figure 4 shows the models' answers to the knowledge-based questions. MLeVLM can effectively give the treatment suggestions, while the other models can't. More case studies in the Appendix G.2 demonstrate that our model can have greater multi-level capabilities.

Case Study on Reasoning Capability. We also conduct the case study on reasoning capability. Plot (c) of Figure 4 shows the reasoning capability of the models. It can be seen that LLaVA-Med and Med-flamingo have the correct answers but lack the correct reasoning process. MLeVLM gives a detailed and reliable progressive reasoning process while ensuring the correct answer. More examples in the Appendix G.2 can demonstrate the strong reasoning capability of MLeVLM compared to the existing medical MLLMs.

7 Conclusion

In this paper, we summarize the multi-level progressive capabilities of medical visual question answering based on real scenarios for the first time. We propose MLeVLM, optimized in terms of both data and architecture. Extensive experiments based on our manually constructed multi-level evaluation benchmark, MLeBench, demonstrate that MLeVLM is the best performing medical MLLM at present. Meanwhile, case studies and human evaluation experiments demonstrate that MLeVLM exhibits great progressive reasoning capability.

Limitations

For the multi-level question-answer generation approach used in constructing the MLe-VQA, it is capable of generating only the questions on the basis of the image captions. Consequently, its scope is bound by the given captions, limiting the information available for question and answer generation. Additionally, despite that we have carefully crafted prompts and in-context examples, GPT-4 may prone to language hallucinations, therefore it might generate incorrect answers. Generally, more trustworthy language models are desired for self-instruct data generation.

Ethics Statement

MLeVLM is built upon a Large Language Model, inheriting the original language hallucinations of the LLM, e.g., it may produce harmful and counterfactual responses. Moreover, machines are imperfect, so a potential risk is that the model may misinterpret user input or make inaccurate predictions. In high-stakes medical environments, such errors could prove harmful or even dangerous. Researchers and developers should be aware of the potential harms that may arise from the use and misuse of MLLMs in medical settings, and should implement both automated (e.g., setting stringent thresholds for diagnostic suggestions) and human (e.g., training to ensure staff awareness of potential system fallibilities) safeguards. We hereby declare that the MLe-VQA dataset we have released is made available solely for research purposes. Furthermore, our data collection method align with the terms of use and adhere to the intellectual property and privacy rights of the original authors.

Acknowledgement

We thank all reviewers for their insightful comments and suggestions. This paper was supported by National Key R&D Program of (2021YFF1201100), Beijing Nova Program and Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences, CI2023C062YLL.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. 2023. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures.

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

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

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023d. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023e. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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)*, pages 1650–1654. IEEE.

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

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*, 35:2507–2521.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pages 13937–13949. Curran Associates, Inc.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

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. In *Findings of EMNLP*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2023a. Towards generalist biomedical ai.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023b. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qi Wu, Peng Wang, Xin Wang, Xiaodong He, and Wenwu Zhu. 2022. Medical vqa. In *Visual Question Answering: From Theory to Application*, pages 165–176. Springer.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*.

Juexiao Zhou, Xiuying Chen, and Xin Gao. 2023a. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *arXiv preprint arXiv:2306.10765*.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. 2023b. Skingpt-4: an interactive dermatology diagnostic system with visual large language model.

A More Related Work

A.1 Construction of the MLLM instruction dataset

A series of researches further enhance MLLMs by focusing on data quality and data diversity in the pre-training and fine-tuning phases. LLaVA inputs captions and bounding boxes into GPT-4 to construct the instruction dataset. Instead of inputting a picture, this method only inputs human-labeled text into the model and lets it imagine that it has seen the picture to constrain the final generated instructions and responses to get the LLaVA-Instruct-150K dataset. InstructBLIP (Dai et al., 2023), starting from BLIP-2, further collects more diverse datasets based on the LLaVA-Instruct-150K dataset to produce instruction tuning data. To build a large-scale high-quality image-text pair data, ShareGPT4V (Chen et al., 2023b) passes 100K images through GPT-4 to generate complex captions. A pre-trained caption model is fine-tuned with these high-quality captions, and 1.2 million images are passed through the fine-tuned caption model to generate complex captions in order to build a sizeable ShareGPT4V dataset. Monkey (Li et al., 2023e) incorporates multiple generators to automatically generate multi-level description generation for images.

A.2 Reasoning in MLLM

The success of rationales in unimodal reasoning (Kojima et al., 2022; Rubin et al., 2021; Wei et al., 2022) motivates a growing body of researches to leverage rationales to enhance both reasoning capabilities and interpretability in multimodal reasoning. The pioneering work (Lu et al., 2022) first presents ScienceQA, a large-scale multimodal Q&A dataset where annotations include both the answers to questions and the corresponding rationals. MMCoT (Zhang et al., 2023b) proposes a two-stage framework that separates rationale generation and answer inference, the first stage is the rationale generation stage, where the model generates the rationals based on the input text and images, and the second stage integrates all the information obtained to generate the final answer. Then, (Wang et al., 2023) employs different strategies for tasks of different difficulty. It utilizes a zero-shot instruction to generate CoT rationals as teaching data for simple tasks. For complex tasks, it decompose them into sub-problems through zero-prompting to construct teaching data and mixing different training data for model fine-tuning. In order to improve the generalization of the generated rational in out-of-distribution data, DDCoT (Zheng et al., 2023) generates rationals by zero-shot prompting to take advantage of the intrinsic generative power of LLM, which prompting LLM to decompose the input problem into sub-problems and acquiring corresponding sub-answers. The obtained information will be integrate into auxiliary information to generate rationales by prompting LLM. The problem statements combined with the rationales are served as inputs for zero-shot prompting or fine-tuning to improve the reasoning abilities.

A.3 Relationship with Previous Works

Compared with previous work (Moor et al., 2023; Li et al., 2023a), our approach is not a direct migration of multimodal large language model from general domain. We focus on improving the multi-level progressive capabilities in medical VQA. On the data side, we carefully construct instruction datasets to allow models with multi-level capabilities and progressive reasoning. On the architecture side, we use multilevel feature alignment module instead of simple linear layers. This ensures that the model can focus on multilevel capabilities from visual to semantic.

B Data Source

B.1 Data Source for Medical Modality Alignment

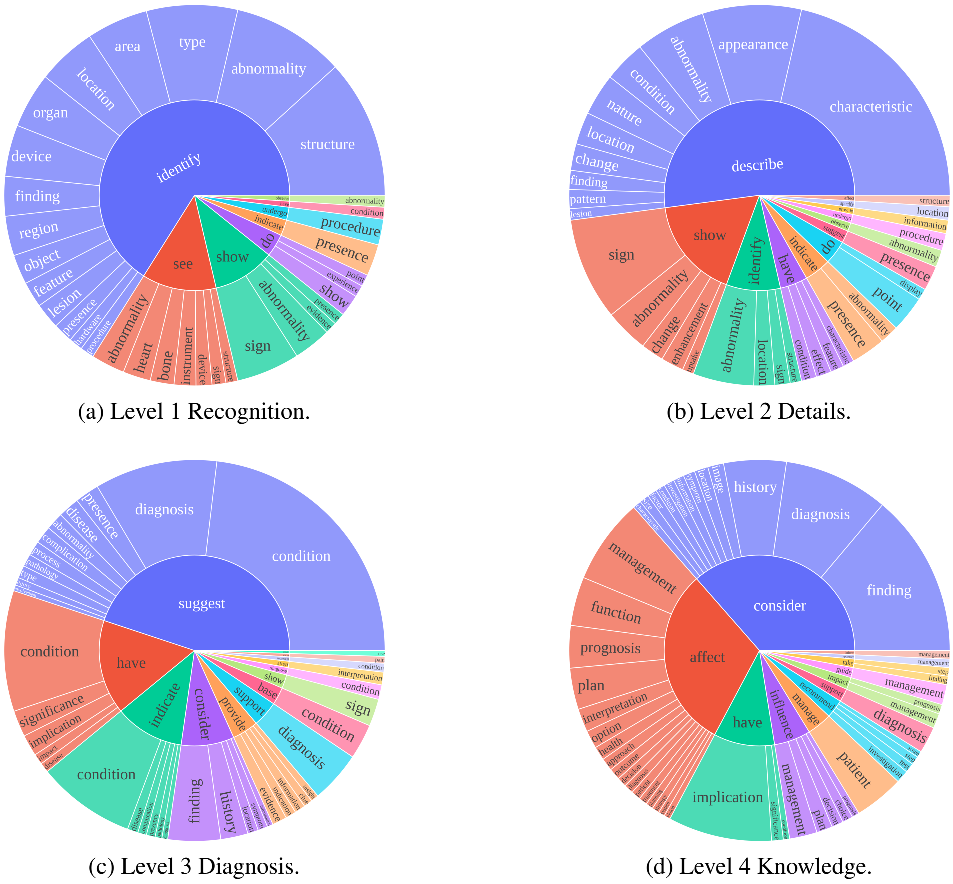
We use multimodal medical data containing three types of tasks as the data source for the first stage of medical modality alignment. It contains a total of 1,710K Q&A samples. The details of the data source are shown in Table 4.

Table 4: Data Source for Medical Modality Alignment.

Task	Datasets	Samples
Diagnosis	MedMINIST (Yang et al., 2023),etc.	762,615
Image Caption	Medicat (Subramanian et al., 2020),etc.	761,641
VQA	PMCVQA (Zhang et al., 2023a),etc,	186,033

B.2 Data Source for MLe-VQA

We select 4 biomedical caption datasets as data sources, including ROCO (Pelka et al., 2018), Medicat (Subramanian et al., 2020), MIMIC-CXR (Johnson et al., 2019) and PMCVQA (Zhang et al.,



(a) Level 1 Recognition.

(b) Level 2 Details.

(c) Level 3 Diagnosis.

(d) Level 4 Knowledge.

Figure 6: The visualization of the root noun-verb pairs for the questions in MLe-VQA.

2023a). We filter high-quality caption data and use them to construct MLe-VQA with multi-level capabilities. The details of the data source are shown in Table 5. The MLe-VQA dataset is split into training, validation, and test splits with 53024, 3514, and 3431 Q&A samples, respectively.

B.3 Data Source for MLe-Bench

We use four common biomedical VQA datasets as data sources, including PathVQA (He et al., 2020), VQA-RAD (Lau et al., 2018), VQA-Med (Ben Abacha et al., 2021) and Slake (Liu et al., 2021). We construct MLe-Bench by careful manual selection and the details are shown in Table 6. The test of reasoning capability is performed on Diagnosis and Knowledge.

C Data Analysis

Figure 6 provides a visualization of the root noun-verb pairs for the questions in MLe-VQA. The visualization results reveal that the questions corresponding to each rank focus well on their own ranks and also demonstrates the diversity of the questions. Figure 7 shows several examples of MLe-VQA. The results of the comparison with the current medical VQA dataset are shown in Table 7. Based on the statistical results it can be seen that our constructed MLeVQA contains the richest modality and the most question-answer pairs. With the rephrasing ability of GPT-4 and the rich data sources, the question types of MLeVQA are more diverse than other medical VQA datasets.

The MLe-VQA contains the following advantages:

1) Laborious manual labelling are avoided by using the latest GPT-4 for automated construction;

2) The construction based on GPT-4 provides a rich diversity and complexity of instructions, which is more conducive to the instruction fine-tuning stage;

3) Obtain more fine-grained medical VQA data through clear definitions of levels and provide reasoning process based on progressive relationships, reducing hallucination generation.

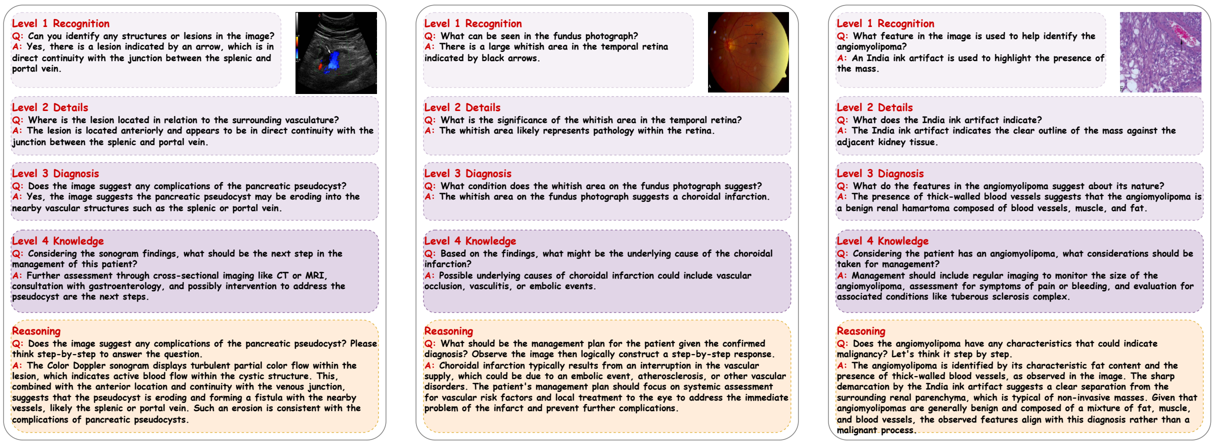


Figure 7: Several examples of MLe-VQA.

Table 5: Data Source for MLe-VQA, including multi-level capabilities sample counts.

Dataset	Recognition	Details	Diagnosis	Knowledge	Reasoning	Total
ROCO (Pelka et al., 2018)	5,242	5,130	4,278	4,839	9,117	28,606
Medicat (Subramanian et al., 2020)	2,037	2,003	1,722	1,897	3,619	11,278
PMCVQA (Zhang et al., 2023a)	2,013	1,988	1,658	1,868	3,526	11,053
MIMIC-CXR (Johnson et al., 2019)	1,614	1,600	1,386	1,523	2,909	9,032
Total	10,906	10,721	9,044	10,127	19,171	59,969

Table 6: Data Source for MLe-Bench, including multi-level capabilities sample counts.

Dataset	Recognition	Details	Diagnosis	Knowledge	Total
Slake	61	92	69	56	278
VQA-RAD	195	88	152	83	518
PathVQA	140	76	156	53	425
VQA-Med	72	92	62	45	271
Total	468	348	439	237	1,492

D More Implementation Details

We instantiate the model with the pre-trained EVA-G[5] (Fang et al., 2023) for visual encoder and Q-Former[6] (Dai et al., 2023) for text encoder. We choose Vicuna-7B[7] (Chiang et al., 2023) as the default large language model. The blocks of context merger are 3 by default. We train all models with 4*A800s. We pretrain our model on Stage I for 1 epoch with a learning rate of 2e-5 and a batch size of 32, and fine-tune on Stage II for 3 epochs with a learning rate of 1e-5 and a batchsize of 4. Finally, we fine-tune the model on Stage III for 2 epoch with a learning rate of 2e-5 and a batchsize of 16. Following LLaVa, we use the Adam optimizer with no weight decay and a cosine learning rate with a warmup ratio of 3%. Details of the training process are shown in Table 8.

[5]https://github.com/baaivision/EVA

[6]https://github.com/salesforce/LAVIS

[7]https://huggingface.co/lmsys/vicuna-7b-v1.5

E Instructions and Prompts

E.1 Instructions for Image Classification

To make the image classification task more suitable for the VQA model, we design 10 question instruction templates and 7 answer templates (Yin et al., 2023) when performing medical modality alignment. The templates are shown in Figure 9. They present the same meaning with natural language variance.

E.2 Instructions for Image Caption

To make the image caption task more suitable for the VQA model, we design 10 instruction templates (Li et al., 2023a) when performing medical modality alignment. The instructions are shown in Figure 10. They present the same meaning with natural language variance.

E.3 Instructions for Reasoning

To enrich the instructions for the reasoning process, we design 10 instructions to inspire the model for reasoning. The instructions are shown in Figure 11. They present the same meaning with natural language variance.

Table 7: Comparisons with Other Medical VQA Datasets.

Dataset	Images	QA Pairs	Number of Images for Different Modalities
Slake	642	14K	CT 282; MRI 181; X-ray 179.
VQA-RAD	315	3.5K	Head axial CT/MRI 104; X-ray 107; Abdominal axial CT 104.
PathVQA	4998	32K	All pathology images.
VQA-Med	2000	36K	All endoscopy images.
MLe-VQA	5352	60K	Xray 1854; CT 1558; MRI 1201; Ultrasound 364; PET 161; Angiogram 78; Pathology 25; Endoscopy 16; Ophthalmic Imaging 15; Others 95.

Table 8: Training Details.

Stage	Para.	Batchsize	LR	Epochs	GPU Hours
Stage I	228 M	32	2e-5	1	144
Stage II	7,072 M	4	1e-5	3	72
Stage III (MLFA)	228 M	16	2e-5	2	24
Stage III (Lora)	388 M	16	2e-5	2	30

E.4 Prompts for MLe-VQA Generation

Figure 12 shows our prompts for generating the MLe-VQA dataset using GPT-4. The prompts include clear definitions of the different levels, reasoning prompts, and in-context learning examples.

E.5 In-context Learning Examples

Figure 13 shows one of our manually labeled in-context learning examples, including the caption for input images, expected multi-level questions and answers, and progressive reasoning processes. The output is in standard JSON format.

E.6 Prompts for GPT-assisted Evaluation

Motivated by (Li et al., 2023a; Liu et al., 2023), we leverage GPT-score to quantify the model response to a question. Specifically, We craft specific evaluation prompts for each capability, informing GPT of the focus of each level, as is shown in Figure 14. We then feed the question, ground truth and the generated model response to the GPT-3. Each response generated by the model receives an evaluation score from GPT, ranging from 0 to 5, where a higher score indicates better overall performance. Each sample is input into GPT three times, and the average score is taken as the final score to ensure stability of the results. Finally, we calculate the average score obtained by each model across all questions as the score for each model.

F More Experimental Results

F.1 Comparison on More Traditional Metrics

We conduct comparative experiments on more traditional metrics, including BLEU and Sentence Similarity. The results are shown in Table 9. It can be seen that MLeVLM fine-tuned with Lora achieves excellent results in evaluation. Among them, it achieves the best similarity in the recognition and diagnosis categories, and the best BLEU in the recognition and knowledge categories. Compared with existing medical MLLMs, the tradional metrics of MLeVLM show strong competitiveness.

F.2 More Analysis of Reasoning Capability

We provide additional analysis on the assessment of the model's reasoning ability. We perform reasoning on the test set of MLe-VQA and automate the evaluation using traditional metrics based on the ground truth generated by GPT-4. The experimental results are shown in Table 10. It can be seen that the reasoning results generated by our model is higher than that of other models on the traditional metrics, proving that it has a more effective reasoning capability. The ablation results leads to the conclusion that the reasoning capability of the model comes from the level instruction-tuning on Stage III.

G More Qualitative Analysis

G.1 Token Visualization

To better demonstrate the effect of our proposed token selector, we conducted visualization experiments. The experimental results are shown in Figure 8. For a given image and question, our attention-based token selector can get the tokens selected by the model as key information. We have labeled the key information with red circles in the figure, and we can find that our token selector can effectively recognize these areas. By selecting these key tokens, the model can effectively answer questions

Table 9: The results of comparison with previous studies on four capabilities. Best and second-best results are shown in **bold** and underline, respectively. BLEU and Sim. refer to BLEU-1 and Sentence Similarity, respectively.

Models	Recognition		Details		Diagnosis		Knowledge	
	BLEU	Sim.	BLEU	Sim.	BLEU	Sim.	BLEU	Sim.
Zero-shot with existing medical MLLM								
Med-flamingo (Moor et al., 2023)	0.264	0.478	0.235	0.497	0.237	0.476	0.186	0.433
LLaVA-Med (Li et al., 2023a)	0.283	0.612	0.308	0.609	0.256	0.668	0.236	0.571
Instruction-tuning with existing MLLM								
LLaMa-adapterV2 (Gao et al., 2023)	0.392	0.634	**0.418**	<u>0.632</u>	**0.391**	0.695	<u>0.324</u>	**0.613**
Minigpt-v2 (Chen et al., 2023a)	0.300	0.605	0.304	0.590	0.263	0.665	0.230	0.565
mPLUG-Owl (Ye et al., 2023)	0.377	0.643	<u>0.414</u>	**0.636**	<u>0.368</u>	<u>0.698</u>	0.310	0.597
Qwen-VL-Chat (Bai et al., 2023)	0.357	0.621	0.407	0.629	0.311	0.663	0.303	0.577
LLaMA-VID (Li et al., 2023d)	0.394	**0.654**	0.392	0.625	0.339	0.680	0.319	<u>0.610</u>
Our methods								
MLeVLM (MLFA)	<u>0.403</u>	<u>0.653</u>	0.408	0.627	0.298	0.626	0.313	0.602
MLeVLM (LoRA)	**0.409**	**0.654**	0.357	0.599	0.365	**0.699**	**0.339**	0.604

Table 10: The results of comparison with other MLLMs on reasoning capability. Best results are shown in **bold**. LT refers to Level-tuning.

Models	Level 3 Reasoning			Level 4 Reasoning		
	BLEU	Rouge	Sim.	BLEU	Rouge	Sim.
LLaVA-Med	0.245	0.223	0.579	0.270	0.222	0.606
LLaMa-adapterV2 w/o LT	0.189	0.205	0.502	0.189	0.191	0.518
LLaMa-adapterV2	0.249	0.223	0.526	0.256	0.218	0.556
QwenVL-Chat w/o LT	0.227	0.223	0.573	0.232	0.216	0.593
QwenVL-Chat	0.282	0.252	0.576	0.274	0.233	0.596
MLeVLM w/o LT	0.234	0.232	0.575	0.266	0.224	0.600
MLeVLM	**0.286**	**0.261**	**0.592**	**0.281**	**0.243**	**0.611**

in the visual recognition and visual details.

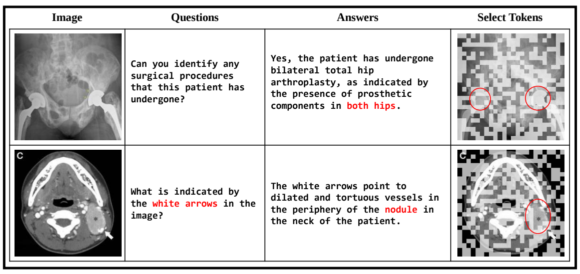


Figure 8: Visualization of the token selector. The key information are labeled with red circles.

G.2 More Cases

We show more cases demonstrating the multi-level capability and progressive reasoning of MLeVLM. The results are shown in Figures 15 and 16. It can be seen that for detail and diagnostic questions, MLeVLM can answer the results correctly. It shows a higher level of detail comprehension and diagnostic capabilities than Med-flamingo and LLaVA-Med. For more modal images, such as colonoscopy and pathology images, MLeVLM can also demonstrate outstanding inference. Other existing multimodal medical models do not succeed in progressive reasoning.

H Details on Human Evaluation

Motivated by (Zheng et al., 2023), for each VQA problem, the evaluators are provided with the image, the question and the correct answer. Evaluators are asked to score the rationales generated by LLaVA-Med (Li et al., 2023a), Med-flamingo(Moor et al., 2023), and our method in five aspects:(1) Relevance: the relevance of the answer to the question; (2) Correctness: the accuracy of reasoning and answer; (3) Completeness: whether the reasoning fully explains the answer; (4) Coherence: the coherence and smoothness of the entire reasoning process; (5) Explainability: the provision of clear and understandable step-by-step reasoning enabling others to comprehend. The rating scale ranges from 0 to 5. We select 50 rationales from Diagnosis and Knowledge at MLe-Bench, respectively, for a total of 100 samples. We recruit clinician evaluators with at least one year of clinical experience as medical professionals for the evaluation. All evaluations are conducted independently by three evaluators. Finally, we average the scores for each aspect of each rationale, resulting in overall scores.

Instructions for Image Classification

- "Can you identify the object in this image?"
- "Based on the image's features, what could be the potential category label for this image?"
- "What label would you assign to this image based on the object's shape and size?"
- "According to the model's prediction, what is the label assigned to this image?"
- "Can you provide the category label for this image based on the object's color and texture?"
- "What label do you think best describes the image's content?"
- "Based on the image's context, what category label would you assign to it?"
- "Can you suggest any alternate labels for this image based on its content and features?"
- "What is the most suitable category label for this image based on its shape, size, and context?"
- "According to the model's classification, what is the category label assigned to this object?"

Answer Templates

- "Upon close inspection of the image, it can be concluded that the majority of the objects in the image fall under the {*C*} category."
- "Through extensive analysis of the image, it can be confidently stated that the image belongs to the {*C*} category."
- "After careful examination of the image, it has been determined that the majority of the objects in the image can be classified as belonging to the {*C*} category."
- "Based on a detailed examination of the image, it can be concluded that the image primarily consists of objects that fall under the {*C*} category."
- "After closely analyzing the image, it has been determined that the main subject of the image belongs to the {*C*} category."
- "Based on a thorough evaluation of the image, it can be confidently stated that the image is dominated by objects that fall under the {*C*} category."
- "The image can be classified as {*C*} based on a close analysis of the objects and their characteristics."

Figure 9: Instruction and anwer templates for image classification.

Instructions for Image Caption

- "Describe the following image in detail."
- "Provide a detailed description of the given image."
- "Give an elaborate explanation of the image you see."
- "Share a comprehensive rundown of the presented image."
- "Offer a thorough analysis of the image."
- "Explain the various aspects of the image before you."
- "Clarify the contents of the displayed image with great detail."
- "Characterize the image using a well-detailed description."
- "Break down the elements of the image in a detailed manner."
- "Walk through the important details of the image."

Figure 10: Instruction templates for image caption.

Instructions for Reasoning

- "Please think step-by-step to answer the question."
- "Let's think it step by step."
- "Let's analyze it piece by piece to answer the question."
- "Let's break down the question and examine the image to respond logically."
- "Observe the image then logically construct a step-by-step response."
- "Reason through the image then answer systematically."
- "Analyze the visual cues in the image then build your answer progressively."
- "Construct your answer in steps."
- "Please formulate your answer in a logical, stepwise fashion."
- "Please systematically articulate your response."

Figure 11: Instruction templates for reasoning.

Prompting GPT-4 to generate medical visual instruction-following data

Please generate questions and corresponding answers based on the image, forming a progressive chain of thought. The thinking chain should include four levels of questions.

Level 1 questions are for testing the ability to observe the image roughly to gain basic information and identify objects.
Level 2 questions, based on the objects identified in Level 1, ask for detailed information in the image, testing the ability to observe visual details.
Level 3 questions are for testing the ability to diagnose diseases.
Level 4 questions test the ability to apply knowledge comprehensively. The questions must be based on this image.

Please provide three parts in your output:
1) The corresponding thought chain questions and answers;
2) A step-by-step detailed reasoning process for the answer to the Level 3 question, incorporating your medical knowledge with findings from Level 1 and Level 2;
3) A step-by-step detailed reasoning process for the answer to the Level 4 question, incorporating your medical knowledge with findings from Level 1、Level 2 and Level

Always pose and answer questions as if you are directly looking at the image and ignore the age information mentioned in the image. Provide two questions for each level. You can only return a dictionary, and the format of the dictionary should strictly follow the example below:

Figure 12: Prompts for MLe-VQA generation.

In-context Learning Example

<Image>: "A female with genital tuberculosis undergoing hysterosalpingography shows beaded appearance in both the right and left fallopian tubes, accompanied by bilateral tubal obstruction (arrows). The uterine cavity contour is mildly irregular."

<Answer>:{"Level 1": [{"Question": "What are the main anatomical structures visible in this image?", "Answer": "The main anatomical structures visible in the image are the uterus, fallopian tubes, and the surrounding pelvic area."}, {"Question": "What imaging technique was used in this image?", "Answer": "This image was obtained through salpingography, which is a specialized X-ray examination."}, {"Question": "What structures do the arrows in the image point to?", "Answer": "The arrows in the image point to the bilateral fallopian tubes."}], "Level 2": [{"Question": "How does the appearance of the fallopian tubes pointed to by the arrows look?", "Answer": "The fallopian tubes exhibit a beaded appearance. The 'beaded' appearance is characterized by uneven expansion and constriction along the length of the fallopian tubes. This is an abnormal appearance, suggesting inflammation and obstruction of the fallopian tubes and implying potential structural issues."}, {"Question": "Is the contour of the uterine cavity normal?", "Answer": "The contour of the uterine cavity is mildly irregular, which may indicate abnormalities within the uterine cavity."}], "Level 3": [{"Question": "Based on this image, what do you think the patient may be suffering from?", "Answer": "The beaded appearance of the fallopian tubes and bilateral fallopian tube blockage suggest the possibility of genital tuberculosis."}], "Level 4": [{"Question": "What conditions other than tuberculosis can lead to similar findings in salpingography?", "Answer": "In addition to tuberculosis, similar findings in salpingography can be caused by conditions such as endometriosis, pelvic inflammatory diseases, and previous surgical scars."}, {"Question": "What are the common symptoms of female genital tuberculosis?", "Answer": "The severity and duration of the condition vary from person to person. Some patients may have no symptoms at all, while others may experience noticeable symptoms, primarily including: (1) Infertility; (2) Menstrual irregularities: early stages may involve heavy menstruation, while later stages may lead to scanty or absent periods, often diagnosed at an advanced stage; (3) Lower abdominal discomfort, worsened during menstruation; (4) During active tuberculosis, general symptoms like fever, night sweats, fatigue, loss of appetite, and weight loss may occur."}], "Inference": [{"Level 3": "From the salpingography image, we see the uterus and fallopian tubes. The fallopian tubes exhibit a beaded appearance and bilateral fallopian tube blockage as indicated by the arrows. This beaded appearance is a characteristic sign of genital tuberculosis, where infection leads to inflammation and scarring of the fallopian tubes. Therefore, the combination of these radiological findings strongly suggests genital tuberculosis as a diagnosis."}, {"Level 4": "Based on the salpingography image, we see the uterus and fallopian tubes exhibiting a beaded appearance and bilateral fallopian tube blockage. This suggests genital tuberculosis, a condition known to cause inflammation and scarring in the female reproductive system. However, other conditions such as endometriosis, pelvic inflammatory diseases, and previous surgical scars can also lead to similar findings. Therefore, while tuberculosis is a possible diagnosis, other possibilities must also be considered. Diagnosis can be assisted by the patient's clinical symptoms,clinical symptoms of female genital tuberculosis include (1) infertility; (2) menstrual irregularities, which may involve heavy menstruation in the early stages and scanty or absent periods in advanced cases; (3) lower abdominal discomfort, aggravated during menstruation; (4) during active tuberculosis, general tuberculosis symptoms like fever, night sweats, fatigue, loss of appetite, and weight loss may occur. The severity and duration of the condition vary from person to person, with some patients having no symptoms and others experiencing noticeable symptoms."}]}

Figure 13: An in-context learning example.

Prompting ChatGPT to evaluate QAs of Level 1

Based on a medical image, the following question is posed:
Question: {question}
The reference answer to the question is: {answer}
The generated answer by AI assistant is: {pred}
Based on the reference answer, evaluate whether the generated answer is consistent with the meaning expressed in the reference answer. The score should be an integer between 0 and 5, where higher scores indicate greater consistency with the reference answer. If the generated answer expresses the meaning contained in the reference answer or is relevant to it, it should be scored appropriately. However, points should be deducted if the generated answer fabricates content related to images or produces content that contradicts medical common sense. Do not provide any other output text or explanations. Only provide a Python dictionary string. For example, your answer should be like this: {'score': }

Prompting ChatGPT to evaluate QAs of Level 2

Based on a medical image, the following question is posed:
Question: {question}
The reference answer to the question is: {answer}
The generated answer by AI assistant is: {pred}
Based on the reference answer, evaluate this generated answer from the following aspects: (1) Accuracy of detail description; (2) Completeness of the overall detail description. Please give an overall score for the generated answer based on the true answer, incorporating your medical knowledge and text comprehension skills, and focusing on the ability to describe details. The score should be an integer between 0 to 5, with higher scores indicating better detail description. In this evaluation, focus on the ability to describe visual details, including but not limited to color, shape, size, number, properties, etc., with more detailed descriptions being better. The priority in the evaluation is accuracy over completeness. Do not provide any other output text or explanations. Only provide a Python dictionary string. For example, your answer should be like this: {'score': }

Prompting ChatGPT to evaluate QAs of Level 3

Based on a medical image, the following question is posed:
Question: {question}
The reference answer to the question is: {answer}
The generated answer by AI assistant is: {pred}
Based on the reference answer, evaluate whether the generated answer is consistent with the meaning expressed in the reference answer. The score should be an integer between 0 and 5, where higher scores indicate greater consistency with the reference answer. If the generated answer expresses the meaning contained in the reference answer or is relevant to it, it should be scored appropriately. However, points should be deducted if the generated answer fabricates content related to images or produces content that contradicts medical common sense. Do not provide any other output text or explanations. Only provide a Python dictionary string. For example, your answer should be like this: {'score': }

Prompting ChatGPT to evaluate QAs of Level 4

Based on a medical image, the following question is posed:
Question: {question}
The reference answer to the question is: {answer}
The generated answer by AI assistant is: {pred}
Based on the reference answer, evaluate this generated answer from the following aspects: (1) Accuracy; (2) Relevance to the question;(3) Helpfulness; (4) Knowledge richness. Note that in this evaluation process, the true answer is only for reference. The generated answer does not have to be exactly the same as the true answer. The generated answer can be richer than the true answer, but a good generated answer should not contradict the facts described in the true answer. You need to use your medical knowledge comprehensively for evaluation and give an overall score for the generated answer. The score is an integer ranging from 0-5. The higher the score, the better the generated answer. Do not provide any other output text or explanations. Only provide a Python dictionary string. For example, your answer should be like this: {'score': }

Figure 14: Prompts for GPT-assisted evaluation of different levels.

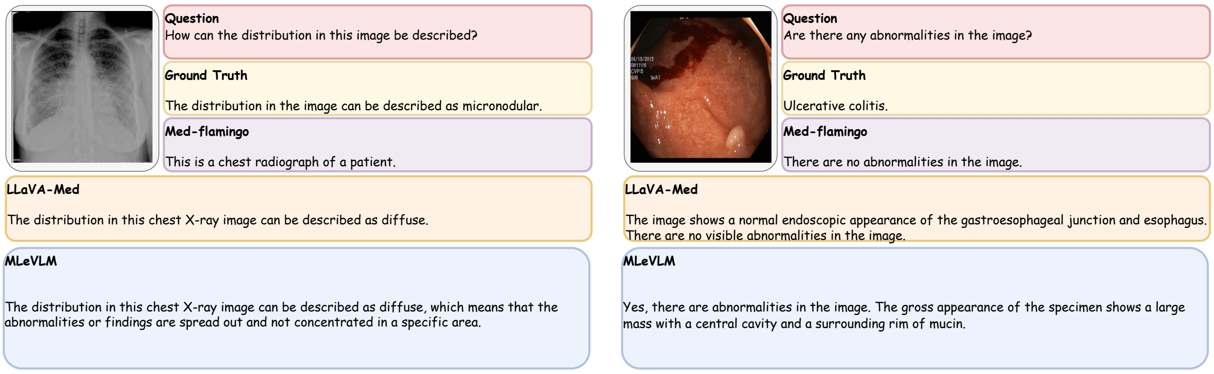


Figure 15: More cases on details and diagnosis.

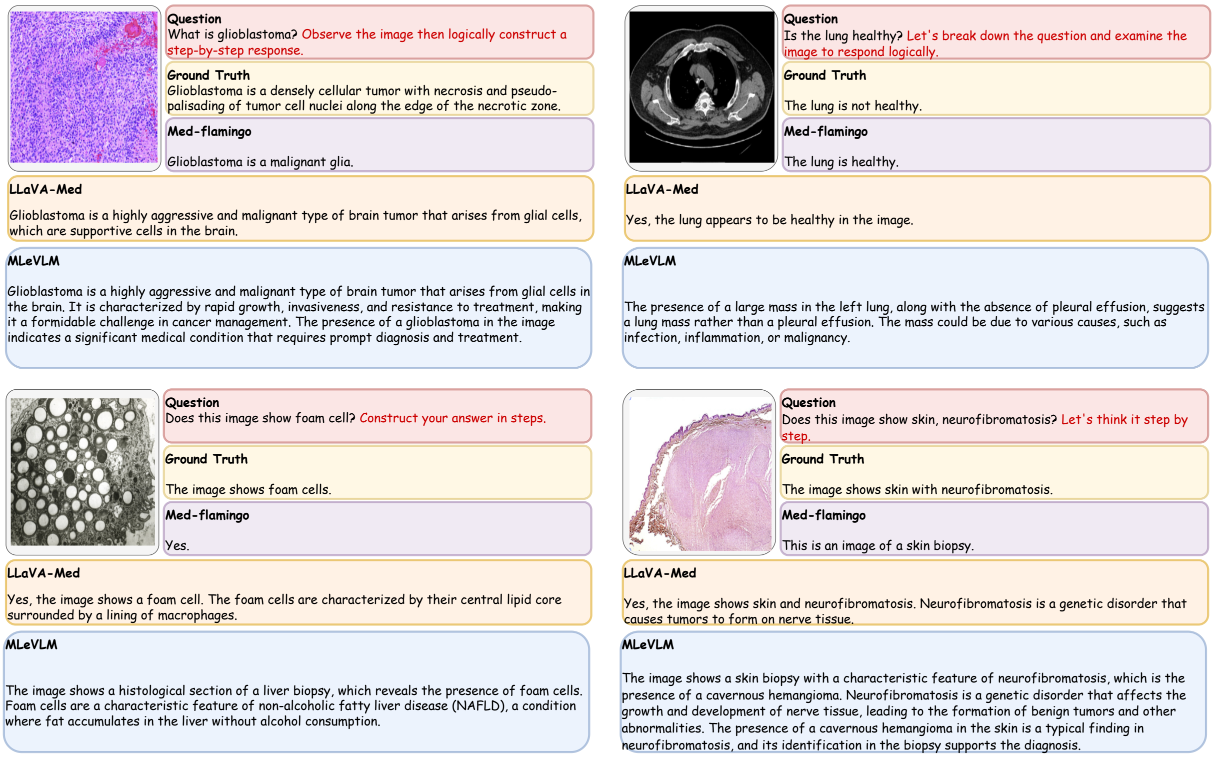


Figure 16: More cases on reasoning capability.