

# Medical Joint Meeting

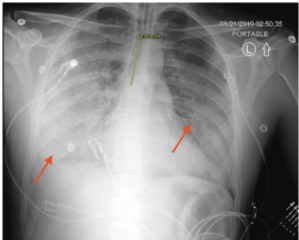
2024.01.05

# LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

Chunyuan Li\*, Cliff Wong\*, Sheng Zhang\*, Naoto Usuyama, Haotian Liu, Jianwei Yang  
Tristan Naumann, Hoifung Poon, Jianfeng Gao

Microsoft  
<https://aka.ms/llava-med>

## Visual input example, Biomedical image from [6]



Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)

Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/>

User	What is shown in this image?
LLaVA-Med	The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.
User	Are there any devices implanted in this patient?
LLaVA-Med	Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.
User	What is shown in this image?
LLaVA	The image shows a ribcage with wires coming out of it, indicating that it is likely a medical illustration or diagram.
User	Are there any devices implanted in this patient?
LLaVA	Yes, there are several devices protruding from the patient's chest, which are likely medical devices used for treatment or monitoring.
User	What is shown in this image?
Language GPT-4	The image shows a chest x-ray taken on Day 2 of a patient's admission after intubation. You can see the endotracheal tube marked by a yellow line, and rapidly worsening ground-glass opacities throughout the lungs. There is also relative subpleural sparing, which is indicated by red arrows.
User	Are there any devices implanted in this patient?
Language GPT-4	Yes, there is an endotracheal tube implanted in the patient, as indicated by the yellow line in the image.

Table 2: Example comparison of medical visual chat and reasoning capabilities. The language-only GPT-4 is considered as the performance upper bound, as the golden captions and inline mentions are fed into GPT-4 as the context, without requiring the model to understand the raw image.

# Background of LLaVA-Med

- Strong zero-shot task performance by **Instruction Tuning**
  - (ICLR 2022) FLAN: **F**inetuned **L**ANguage Models are Zero-Shot Learners
- (ICLR 2023) Visual Instruction Tuning (Generate Instruction from GPT-4)
  - LLaVA : **G**eneral purpose **M**ultimodal conversational assistants. (Biomedical Layperson)
  - LLaVA-Med : Apply multimodal instruction tuning to the biomedical domain. (Biomedical **E**xpert)

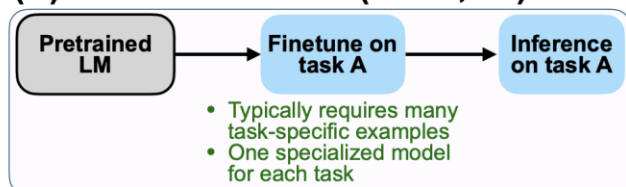
## LM for sentence completion

I went to Jolin's concert last night. I really loved her songs and dancing. It was \_\_\_\_\_

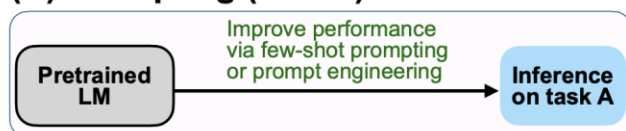
## Detailed task instruction for LM generation

Decide the sentiment of the following sentences:  
I went to Jolin's concert last night. I really loved her songs and dancing.  
OPTIONS: - positive – negative - neutral

### (A) Pretrain–finetune (BERT, T5)

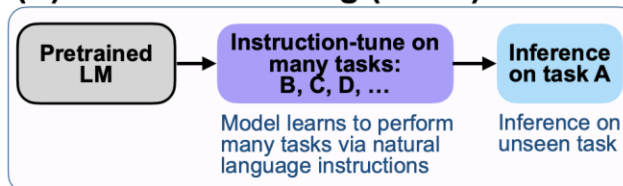


### (B) Prompting (GPT-3)



- Few of Input & Output

### (C) Instruction tuning (FLAN)



- Instruction, Input, Output

→ 任務的詳細敘述

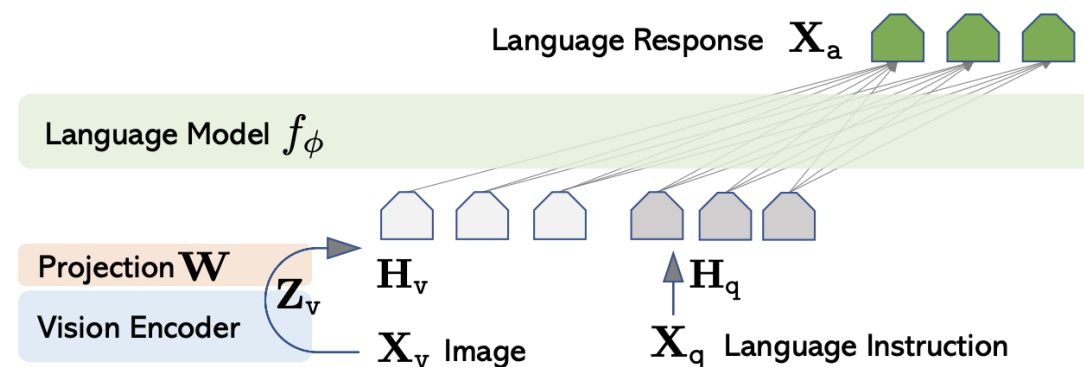


Figure 1: LLaVA network architecture.

# Background of LLaVA-Med

- Vision-Language Model on Biomedical
  - CLIP > BiomedCLIP
    - (Arxiv2023) Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing
  - BiomedCLIP train on **PMC-15M**. (15M text-image pairs from PubMed Central)

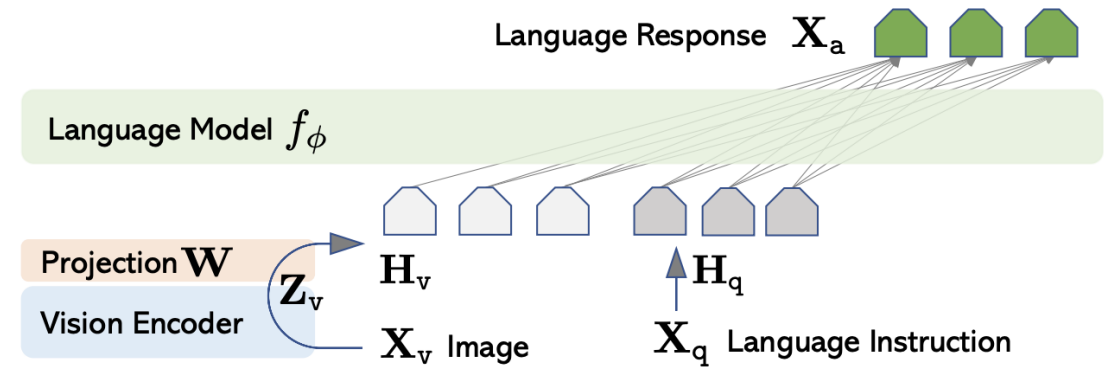


Figure 1: LLaVA network architecture.

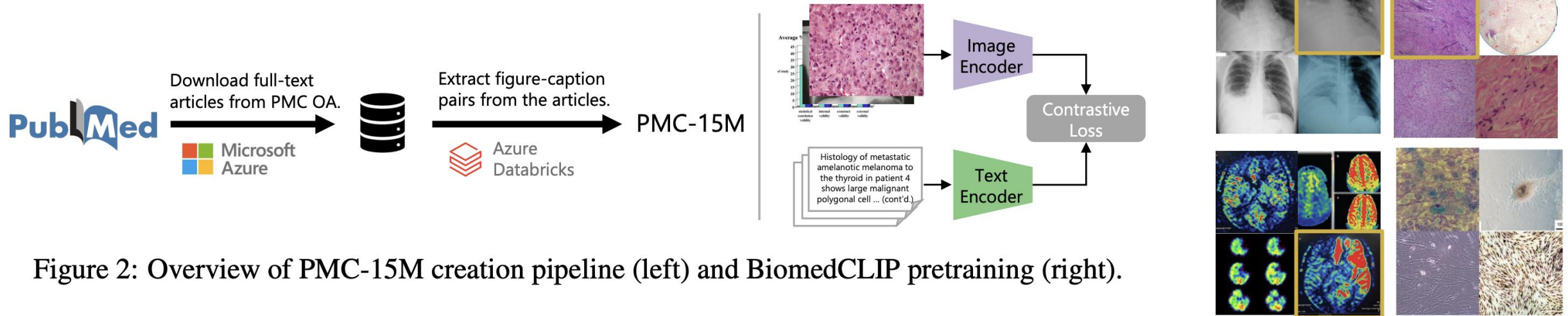


Figure 2: Overview of PMC-15M creation pipeline (left) and BiomedCLIP pretraining (right).

# Related Works

- Visual Med-Alpaca: A Parameter-Efficient Biomedical LLM with Visual Capabilities
  - <https://cambridgeltl.github.io/visual-med-alpaca/>

**Visual Med-Alpaca: Bridging Modalities in Biomedical Language Models**

This is a demo of Visual Med-Alpaca for multi-modal medical foundation model. To use it, simply upload your image and type a question or instruction and click 'submit'.

☐ Input Image

Drop Image Here  
- or -  
Click to Upload

Question/Instruction

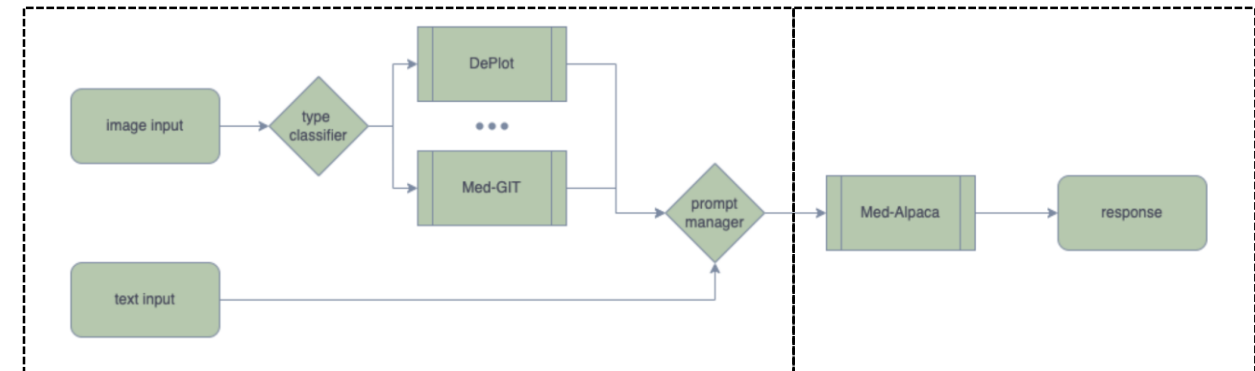
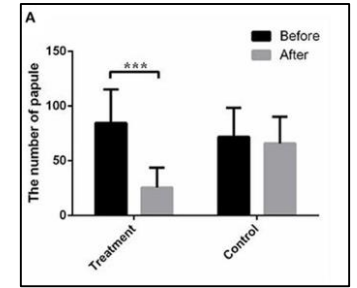
Enter your instruction/question...

LLM

Paste your OpenAI API key (sk-...) and hit Enter (if using OpenAI models, otherwise le...

Submit

Output



Stage1:  
Generate Prompt

Stage2:  
Instruction Tuning

# Method of LLaVA-Med

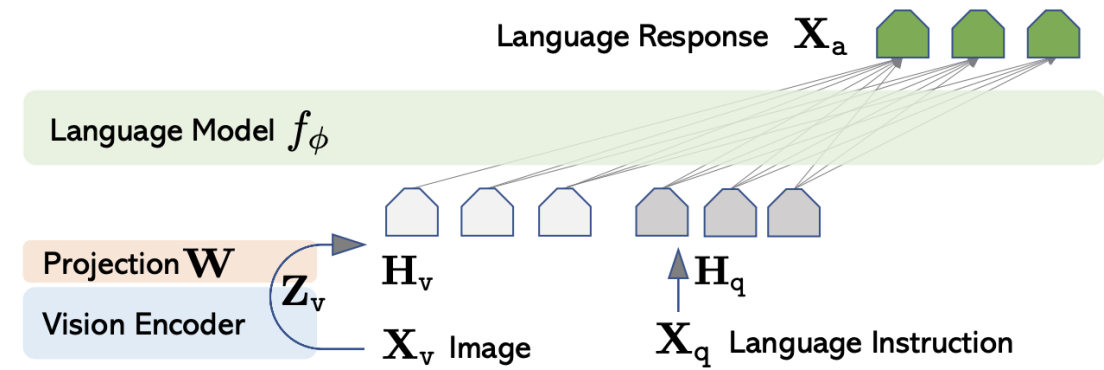


Figure 1: LLaVA network architecture.

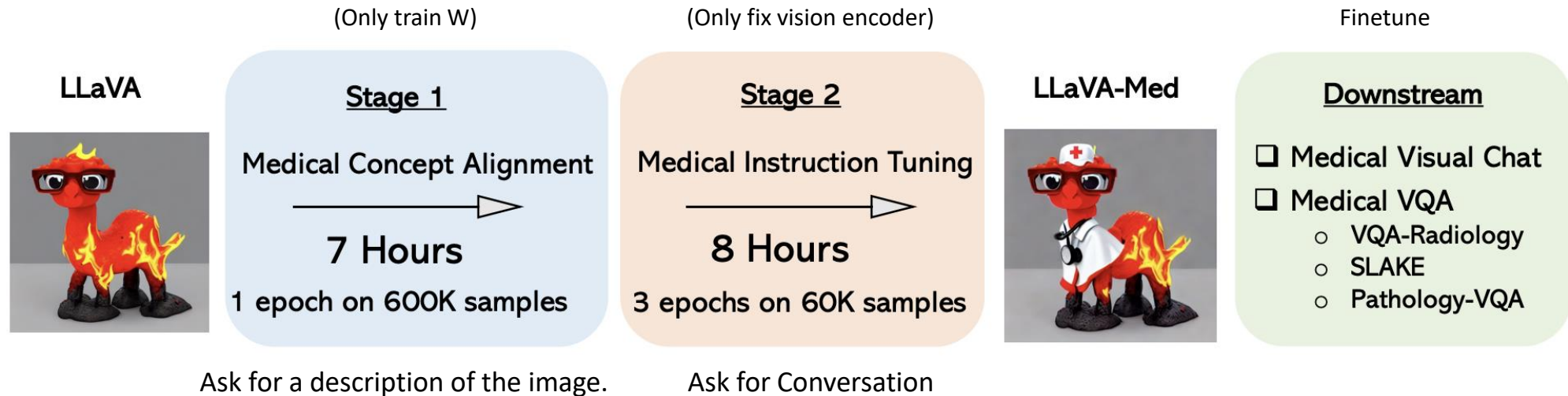


Figure 3: LLaVA-Med was initialized with the general-domain LLaVA and then continuously trained in a curriculum learning fashion (first biomedical concept alignment then full-blown instruction-tuning). We evaluated LLaVA-Med on standard visual conversation and question answering tasks.



# Generate Concept Alignment Data

Human :  $X_q X_v <STOP> \backslash n$  Assistant :  $X_c <STOP> \backslash n$   
 (question). (image). (caption)

**Instructions for brief image description.** The list of instructions used to briefly describe the image content are shown in Table 7. They present the same meaning with natural language variance.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Table 7: The list of instructions for brief image description.

600K data from PMC-15M

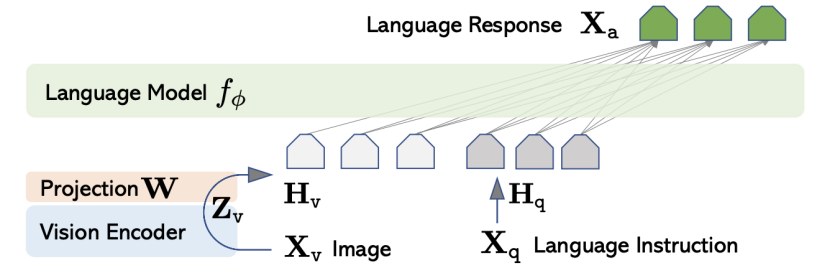
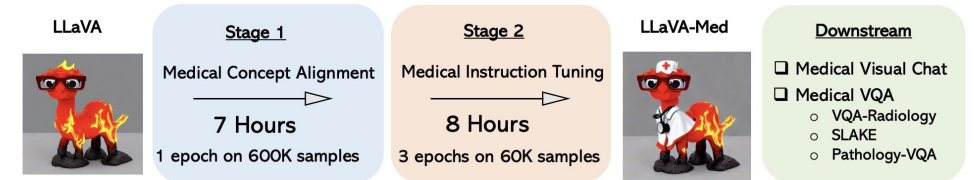


Figure 1: LLaVA network architecture.

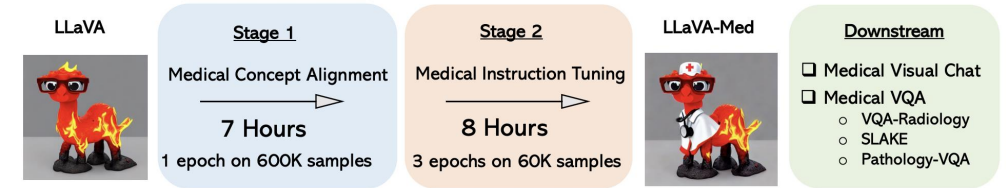


**Instructions for detailed image description.** The list of instructions used to describe the image content in detail are shown in Table 8. They present the same meaning with natural language variance.

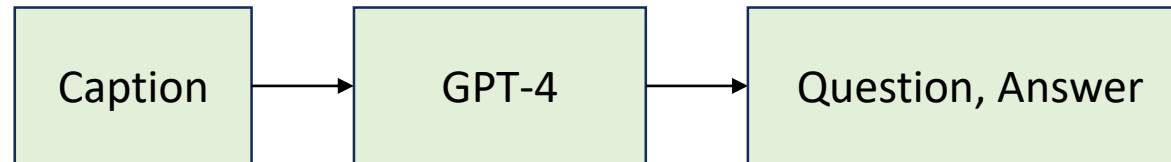
- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before 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"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Table 8: The list of instructions for detailed image description.

# Instruction-Tuning Data



- Multi-round Conversation with GPT-4
- Given an image caption, we design instructions in a prompt that **asks GPT-4 to generate multi-round questions and answers** in a tone as if it could see the image (even though it only has access to the text).





# Experiment of LLaVA-Med

Dataset	VQA-RAD		SLAKE			PathVQA		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2599	858	858
# QA Pairs	1797	451	4919	1053	1061	19,755	6279	6761
# Open	770	179	2976	631	645	9949	3144	3370
# Closed	1027	272	1943	422	416	9806	3135	3391

Table 3: Dataset statistics. For SLAKE, only the English subset is considered for head-to-head comparison with existing methods.

Method	VQA-RAD			SLAKE			PathVQA		
	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
<i>Supervised finetuning results with our own experiment runs</i>									
LLaVA		50.00	65.07		78.18	63.22		7.74	63.20
LLaVA-Med (From LLaVA)		61.52	<b>84.19</b>		83.08	85.34		37.95	<b>91.21</b>
LLaVA-Med (From Vicuna)		64.39	81.98		<b>84.71</b>	83.17		38.87	<b>91.65</b>
LLaVA-Med (BioMed CLIP)		64.75	83.09		<b>87.11</b>	86.78		39.60	<b>91.09</b>
<i>Representative &amp; SoTA methods with numbers reported in the literature</i>									
VL Encoder–Decoder [5]	71.49		82.47				71.49		85.61
Q2ATransformer [26]	79.19		81.20				54.85		88.85
Prefix T. Medical LM [41]				84.30		82.01	40.00		87.00
PubMedCLIP [8]	60.10		80.00	78.40		82.50			
BiomedCLIP [49]	67.60		79.80	82.05		89.70			
M2I2 [22]	66.50		83.50	74.70		91.10	36.30		88.00

(a) Comparison with prior state-of-the-art supervised methods. For open-ended questions, prior methods still formulate the problem as classification among distinct answers in the training set, which may overestimate their generalizability as these datasets are unusual in that the test answers are almost always present in training.


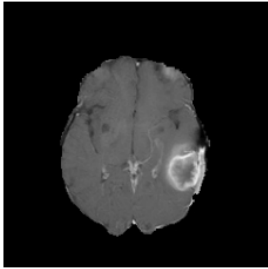
Biomedical image from the SLAKE Bilingual dataset. [23]		
	User LLaVA-Med	这张图片的成像方式是什么？ The image is an abdominal computed tomography (CT) scan, which is a cross-sectional view of the patient's abdomen. The image is taken in the portal phase, which means it is captured after the administration of contrast material to enhance the visibility of certain structures within the abdomen.
	Answer	CT
	User LLaVA	这张图片展示的是核磁共振的哪种类型？ The image shows a typical appearance of a nodular lesion on T1-weighted MR images. The lesion appears hyperintense, which means it is brighter than the surrounding tissue. This can be indicative of various types of lesions, and further evaluation would be needed to determine the specific nature and cause of the lesion.
	Answer	T1

Table 6: Examples on zero-shot on Chinese questions. Though LLaVA-Med training data itself does not include Chinese, it is able to correctly understand the Chinese questions.

**Case Study I: Zero-shot on Chinese Questions.** For the LLaVA-Med trained on 60K-IM data, we provide Chinese questions on SLAKE dataset. Though LLaVA-Med training does not include Chinese instruction-following data, we show in Table 6 that LLaVA-Med is able to correctly understand the Chinese questions and respond the correct answers, probably due to the multilingual knowledge learned in LLaMA/Vicuna. Existing models will fail when zero-shot transfer cross languages.