

488 A Data

489 A.1 Open-sourced Medical Visual Instruction-Following Datasets

490 **Training.** The Stage-1 data follows CC BY NC 4.0 license. As described in Section 3, we create
491 three versions of datasets for our biomedical visual instruction tuning in the 2nd stage.

- 492 • *10K*:
493 [https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/
494 finetune_postprocess_caption_10k.json](https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_10k.json)
- 495 • *60K*:
496 [https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/
497 finetune_postprocess_caption_cleaned_60k.json](https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_cleaned_60k.json)
- 498 • *60K-IM*:
499 [https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/
500 finetune_postprocess_caption_im_cleaned_60k.json](https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_im_cleaned_60k.json)

501 **Evaluation.** As described in Section 5.1, to evaluate the biomedical chat ability, we create an
502 evaluation set.

503 [https://hanoverprod.blob.core.windows.net/public/med_llava/multimodal_chat_
504 eval/qa_50_images.jsonl](https://hanoverprod.blob.core.windows.net/public/med_llava/multimodal_chat_eval/qa_50_images.jsonl)

505 **Images.** The image url paths can be seen in the files:

- 506 • *Training*:
507 [https://hanoverprod.blob.core.windows.net/public/med_llava/images/
508 finetune_image_urls.jsonl](https://hanoverprod.blob.core.windows.net/public/med_llava/images/finetune_image_urls.jsonl)
- 509 • *Evaluation*:
510 [https://hanoverprod.blob.core.windows.net/public/med_llava/images/eval_
511 image_urls.jsonl](https://hanoverprod.blob.core.windows.net/public/med_llava/images/eval_image_urls.jsonl)

512 A.2 Prompts

513 **Instructions for brief image description.** The list of instructions used to briefly describe the image
514 content are shown in Table 6. 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 6: The list of instructions for brief image description.

515 **Instructions for detailed image description.** The list of instructions used to describe the image
516 content in detail are shown in Table 7. 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 7: The list of instructions for detailed image description.

517 **Self-instruct prompts.** The prompts used to generate medical instruction following data are shown
 518 in Figure 4 and Figure 5.

519 **B More Discussions of LLaVA-Med**

520 **B.1 Limitations of LLaVA-Med**

521 **Users of LLaVA-Med.** The penitential users are individuals and professionals within the biomedical
 522 domain who seek assistance in understanding, analyzing, and discussing biomedical images. For
 523 example, (1) Researchers and Scientists: Biomedical researchers working on various topics, such as
 524 CT, Chest X-Ray, MRI, and histology, can use LLaVA-Med to analyze complex biomedical images,
 525 identify patterns, and derive insights from them. For these creative scenarios, models can generalize
 526 easily to provide many new insights based on the large amount of observed samples, while it is
 527 time-consuming for humans to do so. However, models can hallucinate, and humans can select
 528 and revise the model response. (2) Medical Practitioners: Doctors, nurses, and other healthcare
 529 professionals can use LLaVA-Med to improve the working efficiency, as LLaVA-Med can quickly
 530 provide initial answers by understanding diagnostic images, based on which medical practitioners
 531 can improve its factuality without repetitively drafting report from scratch every time. (3) Medical
 532 Students and Educators: LLaVA-Med can serve as an educational tool for medical students and
 533 educators, helping them learn and teach topics related to biomedical images. With the expert approved
 534 simple cases, the AI can help FQA, assisting in explaining concepts, clarifying doubts, and providing
 535 additional context for various imaging techniques and findings.

536 **Limitations of LLaVA-Med.** Precaution is required when utilizing the LLaVA-Med model in
 537 practice: (1) Domain specificity: LLaVA-Med is designed for the biomedical domain, and its
 538 performance may not be as effective in other domains. When testing on other domains, LLaVA-Med
 539 tends to respond with biomedical background knowledge. (2) Reliability: Like other AI models,
 540 LLaVA-Med might inherit biases from the data it was trained on, which could affect its responses.
 541 While LLaVA-Med shows promise in answering open-ended research questions about biomedical
 542 images, its reliability is still subject to the quality and quantity of the training data. The model's
 543 performance on biomedical questions can be improved by fine-tuning, but there is always a possibility
 544 that it may not generalize well to certain types of questions or images not covered in the training
 545 data. Therefore, model hallucination still exists. We strongly suggest users to double-check the

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

```
messages = [ {"role": "system", "content": """"You are an AI assistant specialized in biomedical topics.
```

You are provided with a text description (Figure Caption) of a figure image from a biomedical research paper. In some cases, you may have additional text (Figure Context) that mentions the image. Unfortunately, you don't have access to the actual image.

Below are requirements for generating the questions and answers in the conversation:

- Avoid quoting or referring to specific facts, terms, abbreviations, dates, numbers, or names, as these may reveal the conversation is based on the text information, rather than the image itself. Focus on the visual aspects of the image that can be inferred without the text information.
- Do not use phrases like "mentioned", "caption", "context" in the conversation. Instead, refer to the information as being "in the image."
- Ensure that questions are diverse and cover a range of visual aspects of the image.
- The conversation should include at least 2-3 turns of questions and answers about the visual aspects of the image.
- Answer responsibly, avoiding overconfidence, and do not provide medical advice or diagnostic information. Encourage the user to consult a healthcare professional for advice."""}

```
]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
)
messages.append({"role": "user", "content": query})
```

Figure 4: `messages` we use to prompt GPT-4 to generate medical visual instruction-following data. Manually curated few-shot examples are included in the prompt, where each example has input `sample['context']` and output `sample['response']`. Please see Figure 5 for one of the few-shot examples.

546 responses, and consider them as the preliminary responses that can be revised with expert knowledge.
547 (3) Dependency on input quality: The quality of LLaVA-Med's responses depends on the quality of
548 the input data (biomedical images and captions). Inaccurate or incomplete input data can lead to
549 suboptimal assistance. For example, the current image resolution of the system is 224×224 , which
550 could be too low for the model to see the important details.

551 Without due precautions in practice, the potential negative societal impact of deploying LLaVA-Med
552 can appear. Despite these limitations, LLaVA-Med demonstrates strong potential in assisting with
553 inquiries about biomedical images, encouraging more future research to improve the system.

554 B.2 On the use of GPT-4 API

555 Note that our use of GPT-4 is to convert the text associated with the image into conversational QA
556 format. This is not the typical model distillation as in the language domain, where both teacher and
557 student are text-to-text models. Instead, we trained a (image, text)-to-text model, while the GPT-4 is
558 used as a text-to-text model for data annotation. We discuss both ethical and legal concerns when
559 using GPT-4 or similar LLM to generate self-instructional data.

560 Ethical issues: (1) Accuracy and Misinformation: Generated content may not always be accurate,
561 leading to misinformation being spread, though we have designed a comprehensive filtering script as a
562 post-process to improve the quality. (2) Bias and Fairness: Since we do not have access to the training
563 data of GPT4, the generated instruct data might reflect those biases, reinforcing social or cultural
564 inequalities in the base model training. (3) Deception: In our self-instruct data creation pipeline, the

One of the few-shot examples used in the prompt

Example input:

Figure Caption:

Figure 2: Chest X-ray. Cardiomegaly with diffuse bilateral interstitial infiltrates and a right-sided pleural effusion.

Figure Context:

- His respiratory status continued to worsen and on hospital day number 14 , he required intubation and mechanical ventilation for hypoxemic respiratory failure (Figure 2) .

Example output:

User: Which side of the lung has a pleural effusion?

Assistant: The right side of the lung has a pleural effusion.

User: According to the CT image, does the patient has cardiomegaly?

Assistant: The image appears to be a chest x-ray. However, it indicates that the patient has cardiomegaly, which means an enlarged heart. Additionally, there are diffuse bilateral interstitial infiltrates and a right-sided pleural effusion present in the image.

User: Additional Info: The patient’s respiratory status continued to worsen and on hospital day number 14, requiring intubation and mechanical ventilation for hypoxemic respiratory failure.

Assistant: Thank you for providing additional information about the patient’s condition. Based on the information you provided, it appears that the patient’s respiratory status continued to worsen and on hospital day number 14, they required intubation and mechanical ventilation for hypoxemic respiratory failure. This suggests that the patient’s lung function has significantly declined, and they may be experiencing respiratory distress. It is important for the medical team to closely monitor the patient’s condition and provide appropriate treatment to manage their symptoms and improve their lung function.

Figure 5: One of the few-shot examples used in our prompt to construct medical visual instruction-following data. “Figure Context“ in the input provides additional context by including sentences from the corresponding PubMed paper that mention the figure.

565 GPT-4 API call can be replaced with human annotators, if more budget is available. Since GPT-4’s
566 strong annotation ability is close to humans, if the content is generated without disclosure, it might
567 deceive users into thinking a human produced it. Legal issues: In terms of data usage, we explicitly
568 state that the OpenAI terms should be compiled, and the data can only be used for research purposes.

569 To partially address this concern, we believe that the recently released LLaMA-2-70B-Chat appears
570 to have narrowed the gap. We find that LLaMA-2-70B-Chat can start to follow complex instructions
571 like creating multimodal instructions. However, LLaMA-2-70B-Chat is not correctly following the
572 conversation format. This may be potentially fixed with more sophisticated prompt tuning.

573 **Limitations of data pipeline and the resulting dataset.** Our data pipeline inherits the aforemen-
574 tioned limitations of utilizing GPT-4 API. We considered a comprehensive filtering approach for
575 quality control. Initially, we found that the resulting dataset contains many hallucinated examples.
576 Based on key words of hallucinated examples, we gradually expand our rules to filter out the those

577 examples. Eventually, a comprehensive list of key words are constructed for filtering to increase the
 578 data quality. While the low quality samples probably still exist, we believe our filtering approach is
 579 effective given a limited budget, evidenced by the improved performance using LLaVA-Med on med-
 580 ical VQA datasets. As future directions, one can have experts to revise or filter the generated samples
 581 for higher quality (if more budget is allowed), or get the real-world medical visual conversational
 582 data from clinics.

583 C More Experiment Details

584 C.1 Established Benchmarks

The three established medical VQA datasets are described as below:

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

585

- 586 • *VQA-RAD* [16] contains 3515 QA pairs generated by clinicians and 315 radiology images that
 587 are evenly distributed over the head, chest, and abdomen. Each image is associated with multiple
 588 questions. Questions are categorized into 11 categories: abnormality, attribute, modality, organ
 589 system, color, counting, object/condition presence, size, plane, positional reasoning, and other.
 590 Half of the answers are closed-ended (*i.e.*, yes/no type), while the rest are open-ended with either
 591 one-word or short phrase answers.
- 592 • *SLAKE* [24] is a Semantically-Labeled Knowledge-Enhanced dataset for medical VQA. It
 593 consists of 642 radiology images and over 7000 diverse QA pairs annotated by experienced
 594 physicians, where the questions may involve external medical knowledge (solved by provided
 595 medical knowledge graph), and the images are associated with rich visual annotations, including
 596 semantic segmentation masks and object detection bounding boxes. Besides, SLAKE includes
 597 richer modalities and covers more human body parts than the currently available dataset, including
 598 brain, neck, chest, abdomen, and pelvic cavity. Note SLAKE is bilingual dataset with English
 599 and Chinese. When compared with existing methods, we only consider the English subset.
- 600 • *PathVQA* [13] is a dataset of pathology images. It contains a total of 4998 pathology images
 601 with 32,799 QA pairs. Every image has several questions that relate to multiple aspects such as
 602 location, shape, color, appearance, etc. The questions are categorized into two types, with several
 603 varieties: open-ended questions such as why, what, how, where, *etc.*, and closed-ended questions.

604 **Case Study I: Zero-shot on Chinese Questions.** For the LLaVA-Med trained on 60K-IM data, we
 605 provide Chinese questions on SLAKE dataset. Though LLaVA-Med training does not include Chinese
 606 instruction-following data, we show in Table 12 that LLaVA-Med is able to correctly understand
 607 the Chinese questions and respond the correct answers, probably due to the multilingual knowledge
 608 learned in LLaMA/Vicuna. Existing models will fail when zero-shot transfer cross languages.

609 C.2 Ablation Studies

610 **Impact of Stage-1 training.** We note that it is not always possible to train LLaVA-Med from
 611 LLaVA. For example, we could leverage customized vision encoder (*e.g.*, BioMed CLIP) or LLM
 612 (*e.g.*, Vicuna) to directly train LLaVA-Med using the proposed two-stage training process. The results

613 are reported in Table 9. The customized pre-trained models can provide better performance, *e.g.*,
614 LLaVA-Med trained from BioMed CLIP is better than LLaVA-Med initialized from LLaVA.

615 In the main text, we consider the strategy to train Stage-1 with the linear projection layer only. We
616 now ablate three alternative schemes to study the impact of Stage-1:

- 617 • *Training without Stage-1.* We skip Stage-1, and directly perform medical instruct tuning from
618 LLaVA.
- 619 • *Training Stage-1 with full-model fine-tuning, and including input instruct text.* We keep the same
620 Stage-1 data, but tune the full LLM weights and the linear projection layer.
- 621 • *Training Stage-1 with full-model fine-tuning, and removing input instruct text.* For Stage-1 data,
622 we only consider images as the input, and remove the description-related instruct in Table 6 and
623 Table 7. We tune the full LLM weights and the linear projection layer.

624 The results are reported in Table 10 (a). It yields higher average performance at the early epochs
625 such as epoch 1, when training LLaVA-Med from LLaVA using Stage-2 only, without Stage-1.
626 As the training continues to epoch 3 or more, all training methods perform similarly measured by
627 the average scores. However, training with Stage-1 consistently provides higher performance than
628 training without Stage-1 on the PathVQA dataset (see the comparisons in green cells), which indicates
629 the Stage-1 can benefit certain biomedical domains, when related additional knowledge is learned.
630 Removing the instruct text in Stage-1 that concentrates image description generally improves the
631 performance. This is because LLaVA-Med can smoothly transfer the knowledge of LLaVA in dealing
632 with diverse instruct, without over-fitting to complete the image description tasks.

633 Our suggestions on the necessity of Stage-1 training are (i) If LLaVA-Med is trained with a cus-
634 tomized vision encoder or LLM that are not included in LLaVA (*i.e.*, no LLaVA checkpoint is
635 available), Stage-1 is critical in aligning the multimodal feature space, and yield good performance.
636 (ii) If LLaVA-Med is trained by initializing from LLaVA, the Stage-1 training is optional. In this
637 case, it is more cost-efficient to skip Stage-1 and train Stage-2 only, which can quickly provide good
638 performance on the vertical domains with less cost. However, for scenarios with a large number
639 of in-domain image-text pairs that pre-trained LLaVA does not have much related knowledge, we
640 suggest adding the Stage-1 training on the in-domain pairs: The best strategy in this case is full-model
641 fine-tuning of the LLM, and removing the instruction text of describing the image.

642 **Impact of experiment variance.** In Table 10 (b), we reported multiple experiment run of the same
643 configuration for above Stage-1 training schemes. It turns out the standard derivation of average score
644 is very small. This statistical stability suggest we could use one single run to represent the given
645 experimental configurations. Given the large number of ablation experiments we have performed in
646 this paper, we choose to run the job once for most experiments.

647 **Quality-cost trade-off.** In Table 11 (a), increasing the number of instruct tuning epochs does not
648 improve zero-shot medical VQA performance. Increasing the data size from 10K to 60K improves
649 the average performance by an absolute 1.65% gain, but training cost increases more than four times.
650 The performance gain is limited compared with the additional cost. That’s why we stop further
651 scaling up data size. Instead, we switch to improve the data quality. By comparing 60K-IM and
652 60K datasets, with almost the same training cost, the performance is increased by an absolute 0.86%
653 gain. To achieve a quality-cost trade-off, We suggest more effort devoted to improving the instruction
654 data quality rather than quantity. In Table 11 (b), for fine-tuning stage, we increase the number of
655 fine-tuning epochs on the 60K-IM instruction dataset, and find that the best trade-off is 9 epochs.

656 C.3 More LLaVA-Med Biomedical Chat Results

657 We show more multimodal conversation examples in Table 13, 14, 15.

LLaVA-Med Model Variants				VQA-RAD		SLAKE		PathVQA		Average	
Instruct	Stage 1	Stage 2	FT	Open	Closed	Open	Closed	Open	Closed		
<i>CLIP Vision Encoder [37], 7B Language Model from LLaVA</i>											
LLaVA Init.	0	3	0	0	15.33	15.44	23.61	15.38	6.35	14.74	15.14
	60K-IM	0	1	0	28.93	54.41	39.96	55.29	11.56	54.26	40.73
	60K-IM	0	3	0	31.03	61.76	39.16	55.77	11.43	55.71	42.48
	60K-IM	0	6	0	27.44	59.93	36.35	60.34	11.97	59.42	42.57
	60K-IM	1	3	0	28.23	61.40	39.17	52.16	12.30	54.05	41.22
	60K-IM	1	3	9	66.26	80.88	82.30	84.86	37.59	91.54	73.90
<i>CLIP Vision Encoder [37], 7B Vicuna Language Model</i>											
No LLaVA Init.	60K-IM	1	0	0	16.15	21.32	21.96	15.14	8.07	19.49	17.02
	60K-IM	1	0	9	59.35	76.84	82.74	82.45	38.26	91.42	71.84
	60K-IM	1	3	0	31.71	59.93	38.06	50.96	11.11	49.34	40.18
	60K-IM	1	3	9	64.39	81.99	84.83	83.65	37.76	91.65	74.05
<i>BioMed CLIP Vision Encoder [50], 7B Vicuna Language Model</i>											
No LLaVA Init.	60K-IM	1	3	0	37.84	60.66	39.73	54.33	11.65	49.07	42.21
	60K-IM	1	3	9	64.75	83.09	87.11	86.78	39.60	91.09	75.40
LLaVA	0	0	0	0	20.74	59.19	26.82	50.24	8.74	45.65	35.23

Table 9: Ablation studies of initializing from LLaVA.

LLaVA-Med Model Variants				VQA-RAD		SLAKE		PathVQA		Average
Instruct	Stage 1	Stage 2	FT	Open	Closed	Open	Closed	Open	Closed	
<i>Training Stage-1 with the Linear Projection Layer Only</i>										
10K	1	0	0	15.27	12.50	18.55	13.46	6.26	13.51	13.26
10K	1	3	0	25.79	57.35	31.50	51.68	8.49	59.66	39.08
<i>Training without Stage-1</i>										
10K	0	0	0	20.74	59.19	26.82	50.24	8.74	45.65	35.23
10K	0	1	0	23.61	58.46	36.21	55.05	8.33	56.56	39.70
10K	0	3	0	23.28	62.13	36.50	56.01	7.73	58.12	40.63
10K	0	6	0	28.83	65.81	36.94	60.10	7.87	59.69	43.20
<i>Training Stage-1 with Full-model Fine-tuning, and Including Input Instruct Text</i>										
10K	1	0	0	16.42	13.60	26.16	18.75	8.75	19.52	17.20
10K	1	1	0	28.15	43.01	34.25	41.35	9.29	44.56	33.44
10K	1	3	0	26.23	52.33	37.62	52.48	9.38	58.88	39.49
10K	1	6	0	23.55	56.62	35.58	58.41	9.58	65.56	41.55
<i>Training Stage-1 with Full-model Fine-tuning, and Removing Input Instruct Text</i>										
10K	1	0	0	13.91	7.72	21.65	12.02	6.22	12.56	12.35
10K	1	1	0	27.09	50.00	37.60	53.37	9.26	53.32	38.44
10K	1	3	0	26.88	56.16	35.23	55.78	9.39	63.27	40.97
10K	1	6	0	27.66	61.76	34.43	59.38	9.56	68.03	43.47

(a) The impact of Stage-1 training. All jobs are initialized with LLaVA. It yields higher average performance at the early epochs such as epoch 1, when training LLaVA-Med from LLaVA using Stage-2 only, without Stage-1. As the training continues to epoch 3 or more, all training methods perform similarly measured by the average scores. However, training with Stage-1 consistently provides higher performance than training without Stage-1 on the PathVQA dataset (see the comparisons in green cells), which indicates the knowledge learned in Stage-1 can benefit certain biomedical domains, when related domain data is added.

Jobs	VQA-RAD		SLAKE		PathVQA		Average
	Open	Closed	Open	Closed	Open	Closed	
<i>Training Stage 1 with Full-model Fine-tuning</i>							
Run 1	24.11	54.41	37.73	53.37	9.54	58.09	39.54
Run 2	27.22	53.68	35.68	51.92	9.27	59.10	39.48
Run 3	27.37	48.90	39.45	52.16	9.33	59.45	39.44
Average	26.23	52.33	37.62	52.48	9.38	58.88	39.49±0.10
<i>Training Stage 1 with Full-model Fine-tuning, and Removing Input Text</i>							
Run 1	24.55	59.56	35.27	56.01	9.16	62.22	41.01
Run 2	28.25	53.68	36.02	52.88	9.87	62.72	40.57
Run 3	26.83	56.25	34.40	57.45	9.13	63.88	41.32
Average	26.88	56.16	35.23	55.78	9.39	63.27	40.97±0.56

(b) The ablation study with averaged results from multiple experimental runs. This is experiment setting that initializing from Stage-1 from different schemes, and running multiple jobs of training with 10K instruct samples for 3 epochs. The standard derivation is small, indicating that a single run can represent the experiment configuration setting.

Table 10: Ablation studies of Stage-1 training and experience variance. Zero-shot results on three established biomedical VQA datasets are reported.

Instruct Dataset	10K	60K		60K-IM	
# Epochs in Stage-2	1	1	3	1	3
Performance	39.08	40.73	40.72	41.59	41.22
Training Hours	0.6	2.6	8.0	2.6	8.0

(a) The performance of Stage-2 to tune LLaVA-Med with varied epochs on different instruct datasets.

# Fine-tuning Epochs	0	1	3	9	15	18
Performance	41.22	46.73	65.30	73.90	73.88	73.22
Training Hours	0	0.63	1.37	2.74	4.25	4.93

(b) The performance to fine-tune LLaVA-Med with varied epochs.

Table 11: The quality-cost trade-off. Training hours are reported with batch size 128 on eight A100 GPUs.

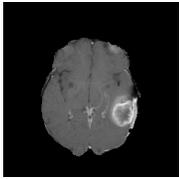
Biomedical image from the SLAKE Bilingual dataset. [24]		
	User	这张图片的成像方式是什么?
	LLaVA-Med	The image is an abdominal computed tomography (CT) scan, which is a cross-sectional view of the patient's abdomen. The image is taken in the portal phase, which means it is captured after the administration of contrast material to enhance the visibility of certain structures within the abdomen.
	Answer	CT
	User	这张图片展示的是核磁共振的哪种类型?
	LLaVA	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 12: 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.

Visual input example, Biomedical image:	
	
User	What is this image about?
LLaVA-Med	The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.
User	What is abnormal about this image?
LLaVA-Med	The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.
User	What is this image about?
LLaVA	The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. The organs are highlighted in different colors , making it easy to identify each organ
User	What is abnormal about this image?
LLaVA	As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...

Table 13: Example comparison of medical visual chat and reasoning capabilities.

Visual input example, Biomedical image:



User	What is the white arrow pointing to?
LLaVA-Med	The white arrow in the MRI image is pointing to a cystic lesion in the pancreatic head.
User	Could this be a mucinous cystic neoplasm?
LLaVA-Med	Yes, it is possible that the cystic lesion in the pancreatic head could be a mucinous cystic neoplasm. Mucinous cystic neoplasms are fluid-filled cysts that can develop in the pancreas. They are usually benign, but in some cases, they can be malignant. Further evaluation and tests would be needed to determine the exact nature of the lesion and the appropriate course of action.
User	Describe the appearance of the mucinous cystic neoplasm
LLaVA-Med	A mucinous cystic neoplasm typically appears as a well-defined, round or oval-shaped cyst with a smooth outer wall. The interior of the cyst is filled with a thick, gel-like fluid called mucin, which gives the cyst its characteristic appearance. The size of the cyst can vary, and in some cases, they can be quite large. It is important to note that the appearance of a mucinous cystic neoplasm on imaging studies, such as MRI or CT scans, can be quite similar to other types of cysts or lesions, so further diagnostic tests and evaluation by a healthcare professional are necessary to confirm the diagnosis.

Table 14: Example comparison of medical visual chat and reasoning capabilities.

Visual input example, Biomedical image:



User	What is the problem with the stomach in this image?
LLaVA-Med	The chest X-ray shows a dilated stomach, which means that the stomach appears to be enlarged or abnormally expanded. This could be due to various reasons, such as an obstruction, motility issues, or other underlying conditions affecting the stomach.
User	How is it affecting the diaphragm?
LLaVA-Med	The dilated stomach is displacing the diaphragm, which means that it is pushing the diaphragm to the contralateral side. This can potentially affect the normal functioning of the diaphragm and may lead to respiratory issues or other complications.

Table 15: Example comparison of medical visual chat and reasoning capabilities.