A Data

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

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

- 10K: https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_10k.json
- 60K: https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_cleaned_60k.json
- 60K-IM: https://hanoverprod.blob.core.windows.net/public/med_llava/finetune_pmc/finetune_postprocess_caption_im_cleaned_60k.json

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

https://hanoverprod.blob.core.windows.net/public/med_llava/multimodal_chat_eval/qa_50_images.jsonl

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

- Training: https://hanoverprod.blob.core.windows.net/public/med_llava/images/finetune_image_urls.jsonl
- Evaluation: https://hanoverprod.blob.core.windows.net/public/med_llava/images/eval_image_urls.jsonl

A.2 Prompts

Instructions for brief image description. The list of instructions used to briefly describe the image 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.

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

<table>
<thead>
<tr>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Describe the following image in detail&quot;</td>
</tr>
<tr>
<td>&quot;Provide a detailed description of the given image&quot;</td>
</tr>
<tr>
<td>&quot;Give an elaborate explanation of the image you see&quot;</td>
</tr>
<tr>
<td>&quot;Share a comprehensive rundown of the presented image&quot;</td>
</tr>
<tr>
<td>&quot;Offer a thorough analysis of the image&quot;</td>
</tr>
<tr>
<td>&quot;Explain the various aspects of the image before you&quot;</td>
</tr>
<tr>
<td>&quot;Clarify the contents of the displayed image with great detail&quot;</td>
</tr>
<tr>
<td>&quot;Characterize the image using a well-detailed description&quot;</td>
</tr>
<tr>
<td>&quot;Break down the elements of the image in a detailed manner&quot;</td>
</tr>
<tr>
<td>&quot;Walk through the important details of the image&quot;</td>
</tr>
<tr>
<td>&quot;Portray the image with a rich, descriptive narrative&quot;</td>
</tr>
<tr>
<td>&quot;Narrate the contents of the image with precision&quot;</td>
</tr>
<tr>
<td>&quot;Analyze the image in a comprehensive and detailed manner&quot;</td>
</tr>
<tr>
<td>&quot;Illustrate the image through a descriptive explanation&quot;</td>
</tr>
<tr>
<td>&quot;Examine the image closely and share its details&quot;</td>
</tr>
<tr>
<td>&quot;Write an exhaustive depiction of the given image&quot;</td>
</tr>
</tbody>
</table>

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

B  More Discussions of LLaVA-Med

B.1 Limitations of LLaVA-Med

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

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

(3) Dependency on input quality: The quality of LLaVA-Med’s responses depends on the quality of the input data (biomedical images and captions). Inaccurate or incomplete input data can lead to suboptimal assistance. For example, the current image resolution of the system is $224 \times 224$, which could be too low for the model to see the important details.

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

B.2 On the use of GPT-4 API

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

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

GPT-4 API call can be replaced with human annotators, if more budget is available. Since GPT-4’s strong annotation ability is close to humans, if the content is generated without disclosure, it might deceive users into thinking a human produced it. Legal issues: In terms of data usage, we explicitly state that the OpenAI terms should be compiled, and the data can only be used for research purposes. To partially address this concern, we believe that the recently released LLaMA-2-70B-Chat appears to have narrowed the gap. We find that LLaMA-2-70B-Chat can start to follow complex instructions like creating multimodal instructions. However, LLaMA-2-70B-Chat is not correctly following the conversation format. This may be potentially fixed with more sophisticated prompt tuning.

Limitations of data pipeline and the resulting dataset. Our data pipeline inherits the aforementioned limitations of utilizing GPT-4 API. We considered a comprehensive filtering approach for quality control. Initially, we found that the resulting dataset contains many hallucinated examples. Based on key words of hallucinated examples, we gradually expand our rules to filter out the those
examples. Eventually, a comprehensive list of key words are constructed for filtering to increase the
data quality. While the low quality samples probably still exist, we believe our filtering approach is
effective given a limited budget, evidenced by the improved performance using LLaVA-Med on med-
cal VQA datasets. As future directions, one can have experts to revise or filter the generated samples
for higher quality (if more budget is allowed), or get the real-world medical visual conversational
data from clinics.

C More Experiment Details

C.1 Established Benchmarks

The three established medical VQA datasets are described as below:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>VQA-RAD</th>
<th>SLAKE</th>
<th>PathVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td># Images</td>
<td>313</td>
<td>203</td>
<td>450</td>
</tr>
<tr>
<td># QA Pairs</td>
<td>1797</td>
<td>451</td>
<td>4919</td>
</tr>
<tr>
<td># Open</td>
<td>770</td>
<td>179</td>
<td>2976</td>
</tr>
<tr>
<td># Closed</td>
<td>1027</td>
<td>272</td>
<td>1943</td>
</tr>
</tbody>
</table>

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

- **VQA-RAD** [16] contains 3515 QA pairs generated by clinicians and 315 radiology images that
  are evenly distributed over the head, chest, and abdomen. Each image is associated with multiple
  questions. Questions are categorized into 11 categories: abnormality, attribute, modality, organ
  system, color, counting, object/condition presence, size, plane, positional reasoning, and other.
  Half of the answers are closed-ended (i.e., yes/no type), while the rest are open-ended with either
  one-word or short phrase answers.

- **SLAKE** [24] is a Semantically-Labeled Knowledge-Enhanced dataset for medical VQA. It
  consists of 642 radiology images and over 7000 diverse QA pairs annotated by experienced
  physicians, where the questions may involve external medical knowledge (solved by provided
  medical knowledge graph), and the images are associated with rich visual annotations, including
  semantic segmentation masks and object detection bounding boxes. Besides, SLAKE includes
  richer modalities and covers more human body parts than the currently available dataset, including
  brain, neck, chest, abdomen, and pelvic cavity. Note SLAKE is bilingual dataset with English
  and Chinese. When compared with existing methods, we only consider the English subset.

- **PathVQA** [13] is a dataset of pathology images. It contains a total of 4998 pathology images
  with 32,799 QA pairs. Every image has several questions that relate to multiple aspects such as
  location, shape, color, appearance, etc. The questions are categorized into two types, with several
  varieties: open-ended questions such as why, what, how, where, etc., and closed-ended 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 12 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.

C.2 Ablation Studies

Impact of Stage-1 training. We note that it is not always possible to train LLaVA-Med from
LLaVA. For example, we could leverage customized vision encoder (e.g., BioMed CLIP) or LLM
(e.g., Vicuna) to directly train LLaVA-Med using the proposed two-stage training process. The results
are reported in Table 9. The customized pre-trained models can provide better performance, e.g., LLaVA-Med trained from BioMed CLIP is better than LLaVA-Med initialized from LLaVA.

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

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

The results are reported in Table 10 (a). 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 Stage-1 can benefit certain biomedical domains, when related additional knowledge is learned. Removing the instruct text in Stage-1 that concentrates image description generally improves the performance. This is because LLaVA-Med can smoothly transfer the knowledge of LLaVA in dealing with diverse instruct, without over-fitting to complete the image description tasks.

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

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

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

### C.3 More LLaVA-Med Biomedical Chat Results

We show more multimodal conversation examples in Table 13, 14, 15.
<table>
<thead>
<tr>
<th>LLaVA-Med Model Variants</th>
<th>VQA-RAD</th>
<th>SLAKE</th>
<th>PathVQA</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instruct</td>
<td>Stage 1</td>
<td>Stage 2</td>
<td>FT</td>
</tr>
<tr>
<td>CLIP Vision Encoder [37], 7B Language Model from LLaVA</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>CLIP Vision Encoder [37], 7B Vicuna Language Model</td>
<td>60K-IM</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>BioMed CLIP Vision Encoder [50], 7B Vicuna Language Model</td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60K-IM</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>LLaVA Init.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9: Ablation studies of initializing from LLaVA.
### Table 10: Ablation studies of Stage-1 training and experience variance. Zero-shot results on three established biomedical VQA datasets are reported.

(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.

(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.
Instruct Dataset | 10K | 60K | 60K-IM
---|---|---|---
# Epochs in Stage-2 | 1 | 13 | 13
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]

<table>
<thead>
<tr>
<th>User</th>
<th>LLaVA-Med</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>这张图片的成像方式是什么?</td>
<td>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.</td>
</tr>
<tr>
<td>User</td>
<td>这张图片展示的是核磁共振的哪种类型？</td>
<td>The image shows a typical appearance of a nodular lesion on T1-weighted MR images. The lesion appears hyper-intense, 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.</td>
</tr>
<tr>
<td>Answer</td>
<td>CT</td>
<td>T1</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>User</th>
<th>LLaVA-Med</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>What is this image about?</td>
<td>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.</td>
</tr>
<tr>
<td>User</td>
<td>What is abnormal about this image?</td>
<td>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.</td>
</tr>
<tr>
<td>User</td>
<td>What is this image about?</td>
<td>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</td>
</tr>
<tr>
<td>User</td>
<td>What is abnormal about this image?</td>
<td>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...</td>
</tr>
</tbody>
</table>

Table 13: Example comparison of medical visual chat and reasoning capabilities.
**Visual input example, Biomedical image:**

![MRI 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:**

![X-ray 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.**