

GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis

Bo Liu¹, Ke Zou^{2,3}, Liming Zhan¹, Zexin Lu¹, Xiaoyu Dong¹, Yidi Chen⁴, Chengqiang Xie¹,
Jiannong Cao¹, Xiao-Ming Wu^{1*}, Huazhu Fu^{5*}

¹The Hong Kong Polytechnic University, Hong Kong, ²National University of Singapore, Singapore,

³Sichuan University, China, ⁴West China Hospital of Sichuan University, China

⁵Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Abstract

Medical Visual Question Answering (Med-VQA) combines computer vision and natural language processing to automatically answer clinical inquiries about medical images. However, current Med-VQA datasets exhibit two significant limitations: (1) they often lack visual and textual explanations for answers, hindering comprehension for patients and junior doctors; (2) they typically offer a narrow range of question formats, inadequately reflecting the diverse requirements in practical scenarios. These limitations pose significant challenges to the development of a reliable and user-friendly Med-VQA system. To address these challenges, we introduce a large-scale, Groundable, and Explainable Medical VQA benchmark for chest X-ray diagnosis (GEMeX), featuring several innovative components: (1) a multi-modal explainability mechanism that offers detailed visual and textual explanations for each question-answer pair, thereby enhancing answer comprehensibility; (2) four question types—open-ended, closed-ended, single-choice, and multiple-choice—to better reflect practical needs. With 151,025 images and 1,605,575 questions, GEMeX is the currently largest chest X-ray VQA dataset. Evaluation of 12 representative large vision language models (LVLMs) on GEMeX reveals suboptimal performance, underscoring the dataset’s complexity. Meanwhile, we propose a strong model by fine-tuning an existing LVLM on the GEMeX training set. The substantial performance improvement showcases the dataset’s effectiveness. The benchmark is available at www.med-vqa.com/GEMeX.

1. Introduction

Large vision language models (LVLMs) have recently made huge advancements in artificial intelligence [1, 2, 13, 34,

Dataset	VQA Type	Example
	Closed-ended	Closed-ended VQA: Q: Is any devices present within the right atrium? A: No.
	Open-ended	Open-ended VQA: Q: What are all the diseases identifiable within the right hilar structures? A: A small right pleural effusion.
	Closed-ended	Multi-choice VQA: Q: Which regions on the X-ray show signs of abnormalities? C: ["A: Bilateral lungs", "B: Right costophrenic angle", "C: Bilateral lower lung", "D: Cardiac region"] A: ["A", "B", "C", "D"]
	Open-ended	Reason: Abnormalities are seen in the bilateral lungs (blunting), right costophrenic angle (blunting), bilateral lower lung (atelectasis), and cardiac silhouette (enlargement).
	Single-choice	Bounding Box: [[28, 40, 167, 190], [28, 124, 85, 184], [28, 126, 165, 189], [57, 114, 147, 180]]
	Multi-choice	

Figure 1. Our GEMeX stands out from existing medical VQA datasets by providing diverse question types and comprehensive multimodal explanations: textual reasoning and visual grounding.

[38, 52, 55], demonstrating remarkable capabilities in understanding visual content while generating coherent natural language responses. These advancements have driven innovations across various domains [15, 46], with healthcare emerging as an important application. Within this domain, medical visual question answering (Med-VQA) stands out as a crucial task that automatically provides reliable and user-friendly answers [29] to questions about medical images [24], facilitating healthcare professionals in diagnosis, medical education, and clinical decision-making.

To ensure the reliability and user-friendliness of Med-VQA systems, it is crucial to incorporate answer explanations along with a diverse set of question formats. Although significant progress has been made by existing Med-VQA systems [16, 18, 24, 30, 50], none have yet integrated explanations for the answers, especially in terms of the visual aspect. As emphasized by [27], explanations are as essential as the answers themselves in general VQA systems. This holds even stronger in medical VQA, where

* Corresponding author.

Dataset	# Images	# QA Pairs	# Modalities	# Question Types‡	# Groundable	# Explainable
VQA-RAD [24]	0.315K	3.5K	Diverse†	O. & C.	✗	✗
SLAKE [30]	0.642K	14K	Diverse†	O. & C.	✗	✗
OmniMedVQA [18]	118.010K	128K	Diverse†	O. & C. & S.	✗	✗
PMC-VQA [50]	149.075K	227K	Diverse†	O. & C. & S.	✗	✗
VQA-Med [7]	4.5K	4.5K	Diverse†	O. & C.	✗	✗
PathVQA [16]	149K	33K	Pathology	O. & C.	✗	✗
RadGenome-Chest CT [51]	50.188K	1.3M	Chest CT	O. & C.	✓	✗
MIMIC-Diff-VQA [17]	164.324K	700K	Chest X-ray	O. & C.	✗	✗
MIMIC-CXR-VQA [4]	142.797K	377K	Chest X-ray	O. & C.	✗	✗
GEMeX (Ours)	151.025K	1.6M	Chest X-ray	O. & C. & S. & M.	✓	✓ (Vision & Language)

Table 1. Comparison of medical VQA Datasets. † indicates a composition of multiple body parts (e.g., head, chest, abdomen) and various imaging modalities (e.g., CT, MRI, X-ray, pathology). In the # Question Types‡ column, O., C., S., and M. represent “Open-ended”, “Closed-ended”, “Single-choice”, and “Multi-choice”, respectively.

the domain-specific nature of the task amplifies the need for clarity [24, 53]. Additionally, the limited range of question formats, such as the absence of multiple-choice questions, restricts the real-world applicability of medical AI systems.

To tackle the aforementioned limitations, we develop a large-scale, Groundable, and Explainable Medical VQA benchmark for chest X-ray diagnosis (GEMeX). We first undertake a comprehensive data refinement process upon the Chest Imagenet [44]. By collaborating with radiologists, we systematically redefine anatomical regions and establish more precise vision-text correspondence mappings, resulting in accurate region-grounded reports for each X-ray image. Subsequently, we leverage GPT-4o [1] to generate a diverse set of questions based on these grounded reports, covering four categories of varying difficulty levels: open-ended, closed-ended, single-choice, and multiple-choice questions. Each question-answer pair is enriched with explicit reasoning and corresponding visual region annotations, as shown in Figure 1. The resulting dataset comprises 151,025 images and 1,605,575 questions. *Currently, this is the largest VQA dataset for chest X-rays and the first medical VQA dataset that simultaneously includes both textual and visual explanations.*

We evaluate 12 representative LVLMs, including 7 from the general domain (e.g., LLaVA [34], DeepSeek-VL [35], GPT-4o-mini [1]), and 5 from the medical domain (e.g., LLaVA-Med [26], XrayGPT [39], RadFM [42]). The experimental findings underscore the challenging characteristics of our dataset. Additionally, we propose a simple instruction-tuning strategy that derives a task-specific LVLM. The impressive performance improvement highlights the effectiveness of our dataset. Overall, we develop three metrics for measuring the accuracy of model outputs in terms of answers, reasoning, and visual grounding (localization generation). Notably, we apply both semantics-level score and gram-based metrics of natural language generation (e.g., BLEU and ROUGE) for textual parts. Re-

sults indicate that for models without GEMeX fine-tuning, semantics-level scoring is more reliable. After fine-tuning, however, the natural language generation metrics can better reflect the model’s understanding of the dataset.

This manuscript makes the following key contributions:

- We present GEMeX, a large-scale medical VQA dataset for chest X-rays, designed to support diverse question types and provide enhanced explainability for medical VQA systems. To our knowledge, it is the largest chest X-ray VQA dataset and the first Med-VQA dataset to embody the concept of multimodal explainability.
- We systematically benchmark 12 representative LVLMs using GEMeX, introducing multiple evaluation metrics to comprehensively demonstrate the performance of current popular LVLMs on the Med-VQA task.
- Our method shows that integrating precise vision-text explainability enhances the visual reasoning of LVLMs, addressing a key limitation in many models. We emphasize the need for a large-scale, groundable, and explainable VQA benchmark to advance LVLM development and deployment in healthcare.

2. Related Work

2.1. Medical VQA Datasets

In recent years, various datasets have been created to advance medical VQA research, each tackling specific challenges in clinical domains. A detailed comparison to other VQA datasets is in Table 1. Specifically, VQA-RAD [24] is a pioneer dataset that offers over 3,000 question-answer pairs focused on radiology images. SLAKE [30] is the first manually created dataset with over 14,000 QA pairs across CT, MRI, and X-ray images, enabling models to handle complex scenarios by combining visual and textual information. VQA-Med [7] is a key dataset for Med-VQA competitions, with 4,500 radiology images and paired QAs. OmniMedVQA [18] provides more data and imag-

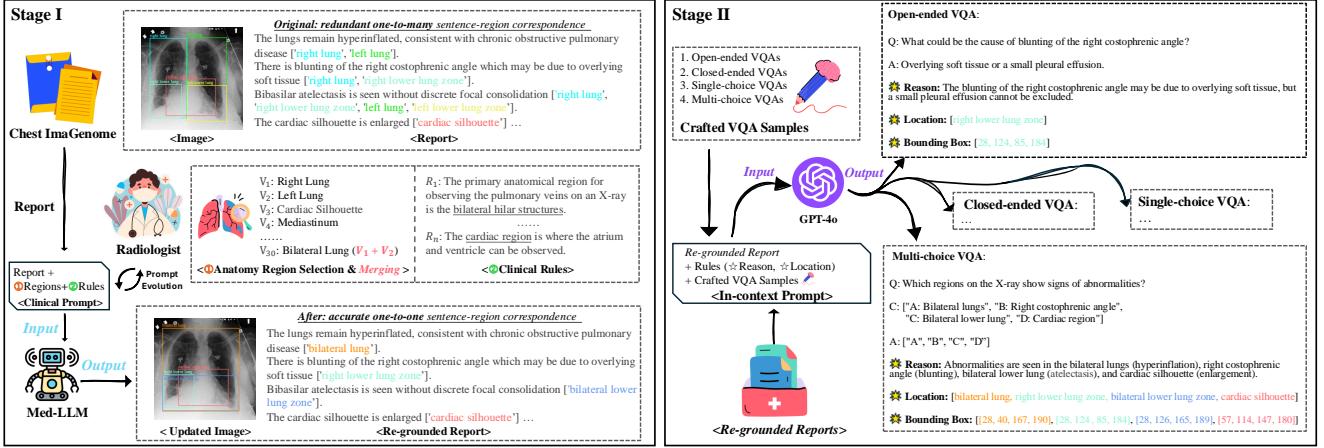


Figure 2. Illustration of the proposed pipeline for constructing our GEMeX, with two main stages. In Stage I (left), medical LLM performs re-grounding on the original reports based on the pathological regions and clinical guidance specified by the radiologists, generating more precise sentence-region correspondence. In Stage II (right), the well-crafted prompt enables GPT-4o to generate a high-quality, large-scale Med-VQA dataset with both textual and visual explanations, leveraging the re-grounded reports from Stage I.

ing modalities, which cover the entire body, to encourage model generalization. PMC-VQA [50] generates VQA data by prompting a large language model to decompose captions of biomedical figures, enabling academic knowledge extraction. PathVQA [16] supplies over 32,000 QA pairs on histopathological images for fine-grained analysis.

For specialized tasks, RadGenome-Chest CT [51] supports chest CT diagnostics, while MIMIC-Diff-VQA [17] emphasizes differential diagnosis reasoning between two X-rays. MIMIC-CXR-VQA [4] expands MIMIC-CXR [21] with diverse question templates to generate thoracic radiology QA pairs, aiding in chest abnormality detection. However, all current datasets lack explainability and diverse question formats. They do not provide detailed visual and textual explanations for answers, which limits their usability for patient and junior doctor comprehension. Additionally, limited question types restrict their ability to simulate the variety of inquiries encountered in practice.

2.2. Medical VQA Methods

Inspired by the advancements in general VQA, medical VQA has gained significant traction as a specialized domain. Due to data limitations, however, most approaches [14, 22, 24, 31, 36, 45] have focused on directly embedding visual and textual information jointly to capture their relationships. With the rise of contrastive language-image pertaining (CLIP) [37], methods [10, 11, 48] start to focus on applying CLIP to Med-VQA. A promising way is to fine-tune CLIP’s joint embeddings to better handle specific medical domains, enhancing the model’s understanding of clinical questions and visual features [28]. Recently, the explosion of large vision language models has further pushed the boundaries of medical domain [3, 25, 26, 39, 42,

56]. Generally, they first pre-train models on a large-scale image-text dataset (like PMC-OA [28], PMC-15M [48]) to map visual features into language model’s embedding space and then further tune with instruction data for medical consultation [3, 34] or disease diagnosis [9, 39, 41, 54]. These models are now leveraged for Med-VQA tasks to provide richer, more context-aware answers, extending beyond simple text-image alignment to incorporate broader knowledge-based reasoning. However, current methods lack explainability, especially in the visual aspect, due to the corresponding absence in current Med-VQA datasets.

3. Construction of GEMeX

We propose a two-stage pipeline for constructing our *GEMeX* dataset, accompanied by an overview in Figure 2.

3.1. Re-grounding Reports

As shown in stage I of Figure 2, we build upon Chest Imagenome [44] to construct our dataset, but with an emphasis on the mapping precision between visual regions and textually described entities. After consulting radiologists, we find that the anatomical descriptions of regions of interest in Imagenome are imprecise and redundant, introducing ambiguity into clinical diagnoses. Specifically, a single sentence can be associated with multiple anatomical regions, e.g., “There is blunting of the right costophrenic angle which may be due to overlying soft tissue” corresponding to “right lung” and “right lower lung zone”. This redundancy poses challenges in training models to precisely visualize grounding. Hence, we perform re-grounding to ensure each sentence is linked to a single, precise region.

Anatomical Region Selection and Merging. In the

original Chest ImaGenome, there are 29 significant pathological regions (with bounding boxes). However, in alignment with radiologists’ practices, our dataset focuses on retaining core regions that are crucial for diagnosing diseases through X-rays, such as the “left lower lung” and “mediastinum”. Less significant or marginal areas are excluded to streamline the diagnostic training process and enhance clinical relevance, like “carina” which is not considered a core region, and “clavicle” which accounts for only about 2% of the total regional frequency. Furthermore, to enhance clarity and ensure that each sentence corresponds to a single pathological region with finer granularity, semantically similar regions are merged. For instance, the “left lower lung zone” and “right lower lung zone” are combined into a “bilateral lower lung zone”. This aligns with conditions like “bibasilar atelectasis”, as illustrated in Stage I of Figure 2, where the condition is described as “Bibasilar atelectasis is seen without discrete focal consolidation”. In total, we define 30 anatomical regions. Detailed transformation from Chest ImaGenome to ours can be found in the Appendix.

Report Re-grounding Using Medical LLM. Then, we utilize OpenBioLLM-70B¹, known for its outstanding performance across various medical NLP tasks, to *re-ground reports by re-selecting a region for each sentence*. To test the effectiveness of the prompt, we begin by randomly selecting 100 pairs from the Chest ImaGenome test set, which includes approximately 367 sentences. Initially, the performance of the LLM is suboptimal due to: (1) inner knowledge about X-ray disease observation areas is not sufficiently precise, as OpenBioLLM is an NLP model that lacks clinical expertise, and (2) when a sentence indeed involves multiple regions that cannot be merged, the model may either output multiple regions or arbitrarily select one.

Iterative Prompt Refinement via Radiologist Feedback: To address these limitations, we employ an iterative approach, gradually incorporating clinical guidance from radiologists and manually-labeled such cases (in-context learning) to refine the prompt, facilitating (1) more accurate region selection; and (2) more effective sentence splitting and rewriting. For example, “The cardiomedastinal silhouette is normal.” is converted into {“The cardiac silhouette is normal.”:“cardiac silhouette”, “The mediastinal silhouette is normal.”:“mediastinum”}, where “cardiomedastinal” corresponds to the “cardiac silhouette” and “mediastinum”. This approach ensures the output clauses align one-to-one with the respective regions. Ultimately, the final prompt is determined with an accuracy of approximately 98.4% on the aforementioned test set, provided in the Appendix.

Figure 2 presents an example of a re-grounded report and its corresponding updated image.

¹<https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B>

3.2. Groundable and Explainable VQA Generation

Although there are many Med-VQA datasets [4, 17] available, some even generated using MIMIC-CXR or Chest ImaGenome, they all have two weaknesses that diminish their practicality: (1) lacking strong explainability, especially the visual guidance, that hinders the user’s understanding; (2) a restricted range of question types, typically confined to open-ended or closed-ended formats with no inclusion of choice-based questions, reducing the flexibility and comprehensiveness. In general, these issues highlight the necessity for more versatile and explainable Med-VQA datasets to enhance their utility in clinical settings.

Data Generation with Quality Control. As shown in Stage II of Figure 2, we generate our VQA dataset based on re-grounded reports. Here, we employ GPT-4o (2024-08-06) [1] as a generator due to its remarkable capabilities in understanding and generating long texts. *We ensure the quality of the generated dataset by:* (1) to ensure a diverse range of question content [30], like “abnormality” and “location”, we identified 7 distinct categories through discussions with radiologists, as illustrated in Figure 4. Then, we manually craft questions covering these 7 types for 30 images, which serve as good demonstrations in the prompt to enhance the generation accuracy and better align with our objectives; (2) we also design specific rules (like not generating questions that need to be answered by comparing two images) to ensure the generated VQAs are answerable; (3) similar to the re-grounding process, we extract 50 samples to observe the effect of the prompt. Only when the overall performance meets our expectations do we proceed with the final dataset generation.

For each image-report sample, we instruct the GPT-4o to generate a total of 11 questions: 3 open-ended VQAs, 2 closed-ended VQAs, 3 single-choice VQAs, and 3 multi-choice VQAs, culminating in approximately 1.6 million VQA pairs. The reports containing less than three sentences will not be used to generate QAs. The generated location is provided as an anatomical region in text format (*e.g.*, “left lower lung zone”). To enable the VQA model to identify the specific location on the image, a post-processing step is required to map the region to bounding box coordinates. Figure 2 illustrates multi-modal explainability, with more examples and the final prompt provided in the Appendix.

4. Statistics of GEMeX

Dataset Split. GEMeX is partitioned in accordance with the distribution of MIMIC-CXR. Specifically, we have 149,535 images with 1,588,975 QA pairs for training, 1,190 images with 12,640 QA pairs for validation, and 300 images with 3,960 QA pairs for testing. Detailed statistics, including question type distribution and the number of bounding boxes, are shown in Table 2.

	Open. (T/B)	Closed. (T/B)	Single. (T/B)	Multi. (T/B)
Train	441,471/466,725	272,323/277,249	441,114/448,810	434,067/861,635
Valid	3,524/3,704	2,145/2,184	3,520/3,599	3,451/6,955
Test	1,144/1,189	543/552	1,300/1,310	973/1,870
Total	446,139/471,618	275,011/279,985	445,934/453,719	438,491/870,460

Table 2. Distribution statistics of question types (T) and number of bounding boxes (B) across data splits.

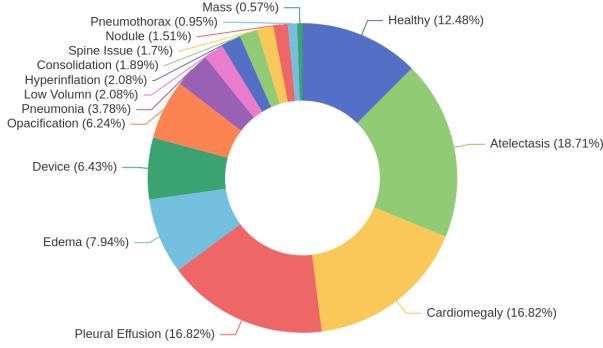


Figure 3. The distribution of normality and abnormality contained in images from the test set of our GEMeX.

Data Quality. To establish a golden test set, we selected 300 images from the MIMIC-CXR’s test set, initially accompanied by 3,291 questions auto-generated by GPT-4o. *Radiologists meticulously reviewed theses questions, correcting around 10 incorrect answers and adjusting 3 inaccurate location annotations. This minimal revision rate demonstrates the high quality of the generated dataset. Additionally, the radiologists contributed approximately 600 new questions, thereby creating a comprehensive golden test set for benchmarking large vision language models.*

Distribution of Clinical Observations. The X-ray image selection for the test set is according to the clinical observations. In Figure 3, we plot the distribution of normality and abnormality contained in images. The original distribution of MIMIC-CXR, characterized by a high proportion of healthy samples, introduces a significant bias affecting model performance [32]. To mitigate this issue, we preserve only a small proportion of healthy samples (around 12.48%) during manual cleaning. Meanwhile, we ensure that clinically important observations occupy a large proportion, such as “atelectasis”, “cardiomegaly”, “edema”, and “pleural effusion” [19]. Additionally, common diseases or observations like “pneumonia”, “opacification”, and “pneumothorax” are also included.

Distribution of Question Content. We show the distribution of the question content categories that GPT-4o itself summarizes according to demonstrations during VQA generation. Figure 4 shows the corresponding results where we can find that “abnormality”, “disease”, and “location”

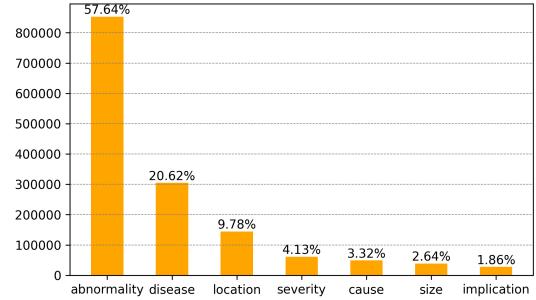


Figure 4. The distribution of question content in our GEMeX.

“account for over 88%, while the remaining categories mainly include “cause”, “size”, “severity”, and “implication”, which highlights the diversity of questions.

More data statistics, such as word cloud and word length distribution, are provided in the Appendix.

5. Evaluation

5.1. Models

A Strong Baseline Fine-Tuned on GEMeX. To validate the effectiveness of the dataset, especially the auto-generated training set, we propose a question-type-aware instruction tuning to fine-tune LLaVA-Med-v1-7B [26] on the training set of GEMeX, termed as LLaVA-Med-GEMeX, serving as a strong baseline. Specifically, for each VQA sample from our GEMeX, we add a type prompt \mathbf{X}_{Type} after the original system prompt and a question $\mathbf{X}_{\text{Question}}$ with its answer $\mathbf{X}_{\text{Answer}}$, textual reason $\mathbf{X}_{\text{Reason}}$, and corresponding visual location $\mathbf{X}_{\text{Location}}$, constructing a single-turn dialogue as in Table 3. Generally, \mathbf{X}_{Type} is “Input a {Type} question, and the assistant will output its answer {Supplement} with a detailed reason and corresponding visual location.” where {Type} refers to “open-ended/closed-ended/single-choice/multi-choice” and {Supplement} is replaced by “none/(yes or no)/(an option)/(some options)”, respectively. Some input samples, as well as the fine-tuning details, are shown in the Appendix.

```
<Ori_System_Prompt>  $\mathbf{X}_{\text{Type}}$  <STOP>
Human: <image>\n $\mathbf{X}_{\text{Question}}$   $\mathbf{X}_{\text{Choices}}$  (if any)
<STOP>
Assistant: <answer> $\mathbf{X}_{\text{Answer}}$  <reason> $\mathbf{X}_{\text{Reason}}$ 
<location> $\mathbf{X}_{\text{Location}}$  <STOP>
```

Table 3. Input format for fine-tuning LLaVA-Med.

Existing LVLMs. Besides fine-tuning a task-oriented model, we perform a zero-shot evaluation on our GEMeX dataset across the other 12 LVLMs, with 7 in the general domain and the other 5 in the medical domain:

Models	Open-ended		Closed-ended			Single-choice			Multi-choice			Avg. \dagger
	AR-score \dagger	V-score	A-score	AR-score \dagger	V-score	A-score	AR-score \dagger	V-score	A-score	AR-score \dagger	V-score	
Random	-	-	48.80	-	-	25.85	-	-	7.50	-	-	-
GPT-4o-mini [1]	97.68	<u>18.05</u>	59.30	71.14	<u>28.64</u>	<u>59.00</u>	<u>77.47</u>	<u>23.62</u>	<u>49.13</u>	<u>82.91</u>	<u>19.19</u>	<u>82.30</u>
LLaVA-v1 [34]	76.14	-	30.76	38.02	-	-	50.47	-	-	66.52	-	57.79
LLaVA-v1.5 [33]	77.62	-	58.93	57.00	-	47.00	57.05	-	-	65.17	-	64.21
Mini-GPT4-v1 [55]	55.32	-	26.33	31.09	-	-	37.63	-	-	46.65	-	42.67
mPLUG-Owl [47]	76.73	-	27.26	36.70	-	32.00	46.89	-	-	67.92	-	57.06
DeepSeek-VL [35]	79.30	11.00	57.10	59.86	8.28	51.69	62.03	8.57	17.99	70.35	12.98	67.89
Qwen-VL-Chat [6]	78.36	3.17	23.02	45.79	12.25	44.69	59.15	16.69	7.30	67.21	2.26	62.63
LLaVA-Med-v1 [26]	90.34	-	62.62	69.91	-	-	61.74	-	-	68.14	-	72.53
LLaVA-Med-v1.5 [26]	94.43	-	<u>71.82</u>	<u>76.54</u>	-	-	66.04	-	-	67.28	-	76.07
MiniGPT-Med [3]	86.12	-	55.24	65.25	-	-	55.61	-	-	64.33	-	67.83
XrayGPT [39]	81.17	-	-	68.17	-	-	48.33	-	-	55.10	-	63.19
RadFM [42]	88.57	-	58.01	67.91	-	-	57.82	-	-	62.41	-	69.18
LLaVA-Med-GEMeX	<u>97.05</u>	51.47	77.35	80.72	53.20	73.08	81.42	54.57	67.42	84.98	47.99	86.04

Table 4. Performance of representative LVLMs on our GEMeX across four different question types. The AR-score combines answer and reasoning to evaluate textual output performance. \dagger denotes the GPTScore value (%). The A-score indicates answer or choice accuracy (%), and the V-score represents mIoU (%). The best results are bolded, and the second-best are underlined in each column.

- **In the General Domain:** we test LLaVA-v1 [34], Mini-GPT4-v1 [55], mPLUG-Owl [47], LLaVA-v1.5 [33], DeepSeek-VL [35], Qwen-VL-Chat [6], and GPT-4o-mini (2024-07-18). Note that we did not test GPT-4o because its safety protection policy prohibits it from analyzing medical images, especially when asking about the condition of patients.
- **In the Medical Domain:** we evaluate LLaVA-Med-v1 [26], LLaVA-Med-v1.5 [26], MiniGPT-Med [3], XrayGPT [39], and RadFM [42]. A detailed introduction can be found in the Appendix.

5.2. Evaluation Metrics

In GEMeX, each question has a corresponding answer, textual reason, and visual location. Ideally, we aim to evaluate all these three aspects with designed metrics as follows:

- **Answer-Reason Score (AR-score):** In reality, most LVLMs struggle to generate accurate outputs in terms of format. This doesn't necessarily mean these models lack the knowledge to answer the questions but rather simply lack the ability to follow instructions properly. To ensure a fair comparison, we introduce the Answer-Reason score (AR-score) as an evaluation metric for the textual output, where the answer and reason parts from each test sample are merged as a reference (ground truth), and the evaluated LVLM's output serve as a candidate. We use GPTScore [26] to calculate the AR-Score from a semantic perspective. Specifically, GPT-4o is leveraged to quantify the correctness by treating the aforementioned reference as a textual response from assistant #1, while the candidate as the response from assistant #2. With both responses, the original question, and the X-ray report, GPT-

4o assesses the accuracy, relevance, and helpfulness of each assistant's answer and provides an overall score on a scale of 1 to 10, where a higher score indicates better performance. We then calculate the relative score using GPT-4o's reference score for normalization. Besides, we also employ common NLG metrics (*e.g.*, BERTScore [49], BLEU, ROUGE) to evaluate AR-score.

- **Answer Score (A-score):** For responses where the model can output specific answers (such as yes/no for closed-ended questions or options for single/multiple choice questions), we calculate the accuracy by comparing with the ground truth. It is worth noting that although some models cannot directly output the answer, we still attempt to match it from their responses.
- **Visual Score (V-score):** For models capable of visual grounding (*i.e.*, outputting visual locations), we calculated mean intersection over union (mIoU) as a measurement. For a VQA case, considering there might be multiple corresponding locations (commonly seen in multi-choice questions), we use the Hungarian algorithm [23] to match the predicted bounding boxes with the actual ones.

5.3. Results and Analysis

Overall Performance. The comprehensive results are shown in Table 4. The first 7 rows indicate the performance of general LVLMs, while the last 6 rows present the results of medical ones and our fine-tuned version of LLaVA-Med-v1 (termed as LLaVA-Med-GEMeX). It can be found that:

- **Most existing LVLMs exhibit weak performance when tested on GEMeX.** The only exception is GPT-4o-mini, which achieves an AR-score above 80 on average across all tasks. When considering specific question types, LLaVA-Med (both versions 1 and 1.5) stands out

Models	Open-ended			Closed-ended			Single-choice			Multi-choice		
	BERTScore	ROUGE-L	BLEU-1	BERTScore	ROUGE-L	BLEU-1	BERTScore	ROUGE-L	BLEU-1	BERTScore	ROUGE-L	BLEU-1
GPT4o-mini [1]	<u>30.43</u>	<u>22.67</u>	<u>18.25</u>	40.02	25.63	19.10	<u>48.34</u>	<u>39.17</u>	<u>30.82</u>	<u>46.58</u>	<u>39.20</u>	<u>28.65</u>
LLaVA-v1 [34]	20.09	15.22	11.57	22.42	13.10	8.01	20.25	14.97	10.61	19.69	17.35	11.15
LLaVA-v1.5 [33]	21.49	16.11	12.20	32.59	15.37	6.69	17.42	17.53	1.49	23.74	21.20	8.95
Mini-GPT4-v1 [55]	15.03	14.66	11.46	13.83	9.65	6.31	6.50	6.79	4.60	5.31	5.79	3.22
mPLUG-Owl [47]	22.52	17.03	13.22	32.23	20.20	13.92	<u>39.64</u>	33.69	30.32	26.09	24.97	16.68
DeepSeek-VL [35]	24.06	18.62	15.94	26.12	23.27	13.83	26.16	30.46	18.79	22.10	27.32	20.74
Qwen-VL-Chat [6]	23.31	18.48	14.63	33.18	22.43	17.19	25.47	22.25	6.95	22.03	22.88	12.26
LLaVA-Med-v1 [26]	25.14	19.63	15.93	38.04	29.08	19.74	34.89	30.10	25.84	28.63	26.51	20.99
LLaVA-Med-v1.5 [26]	26.42	21.38	17.28	<u>44.48</u>	<u>36.73</u>	<u>26.35</u>	36.62	30.32	25.44	28.11	24.49	16.53
MiniGPT-Med [3]	23.47	19.20	16.03	34.31	29.47	19.13	30.11	28.51	22.13	26.51	24.42	15.98
XrayGPT [39]	22.57	18.30	15.73	21.35	14.55	10.17	16.31	12.17	9.23	12.15	10.30	6.22
RadFM [42]	24.96	20.71	17.73	37.43	27.95	20.56	32.30	27.02	24.39	25.81	20.02	13.80
LLaVA-Med-GEMeX	42.69	32.75	25.28	54.44	38.39	33.99	56.35	53.23	47.31	54.95	50.85	43.99

Table 5. Performance of representative LVLMs evaluated using various natural language generation metrics for AR-score, including BERTScore, ROUGE-L, and BLEU-1. The best results are bolded, and the second-best are underlined in each column.

for its strong performance on open-ended questions, scoring above 90 on the AR-score. However, all models show poor results on the other three categories of tasks.

- **When faced with choice-based questions, most models, particularly those in the medical domain, struggle to provide definitive answers, despite their ability to analyze each option.** This difficulty accounts for why many models have an associated AR-score but lack an A-score, highlighting the importance of introducing these types of questions.
- **Powerful LVLMs, such as GPT-4o-mini, often rely on shortcut reasoning rather than real multimodal reasoning.** Although these models can sometimes answer questions to a certain extent (as indicated by the AR-scores), they often fail to accurately achieve visual grounding. This suggests that these models tend to address Med-VQA tasks using shortcut knowledge, such as retrieving information from their pre-training memory, instead of engaging in real multimodal reasoning [8]. However, real multimodal reasoning is essential for the explainability of Med-VQA systems.
- **Through simple question-type-aware instruction tuning, the proposed baseline model achieves a significant performance improvement,** with an approximate **13.5%** increase in average AR-score compared to LLaVA-Med-v1. Notably, it surpasses GPT-4o-mini on most metrics, demonstrating the reliability of the training set. However, a substantial gap remains for practical application, highlighting the challenges associated with GEMeX.

Limitation. Note that the proposed baseline model is inherently task-specific, which may result in reduced accuracy on other tasks or a diminished capacity for conversational engagement. The true potential of our GEMeX lies in its integration into multi-task training frameworks, such as the second training stage of LLaVA-Med. The baseline model primarily serves to demonstrate the dataset’s effec-

tiveness while also providing a robust benchmark.

More Metrics. As mentioned in Section 5.2, we also calculate NLG metrics to measure AR-score. Detailed results are shown in Table 5. Overall, the NLG metrics generally share the same trend as GPTScore (AR-score in Table 4), but there are some minor differences. (1) **High NLG scores do not always correlate with strong model performance**, as seen with mPLUG-Owl compared to LLaVA-v1.5. Essentially, LLaVA-1.5 demonstrates higher performance, such as achieving an answer accuracy rate (A-score) in single-choice tasks that is **15%** higher than that of mPLUG-Owl. However, since LLaVA-v1.5’s output mostly consists of the answer without reason, the shorter output results in a lower NLG score, with its BLEU-1 approximately **28.8%** lower than mPLUG-Owl; (2) **Fine-tuning on GEMeX results in more pronounced improvements in NLG metrics compared to GPTScore.** For example, the fine-tuned baseline model shows only a 3.7% average improvement over GPT-4o-mini on GPTScore, but achieves a 12.1% improvement on average NLG metrics. This significant enhancement better reflects the model’s learning from the dataset. We suggest using GPTScore for models not fine-tuned on GEMeX due to its focus on semantic understanding, while NLG metrics are preferred for fine-tuned models as they better capture alignment with the dataset.

Transferability. To further validate the effectiveness of GEMeX, we assess it from a transfer learning perspective by treating GEMeX as a pre-training dataset and subsequently testing our LLaVA-Med-GEMeX model on other X-ray VQA datasets. Considering potential overlap in data sources, we utilize the CXR portion of the SLAKE test set [30], which includes 505 open-ended questions and 235 closed-ended questions. The zero-shot performance of LLaVA-Med-GEMeX and LLaVA-Med-v1 on this dataset

Challenging examples from GEMeX:

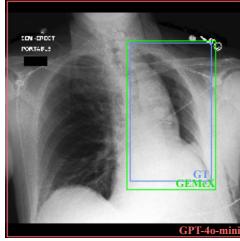
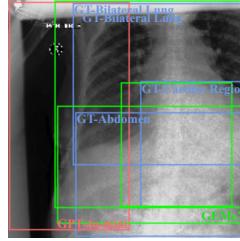
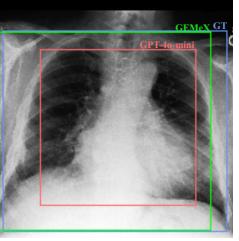
CASE I				CASE II				CASE III				
												
(Bbox predicted by each model)				(Bbox predicted by each model)				(Bbox predicted by each model)				
Question	What might be the cause of the overexpansion of the right lung?			What regions on the X-ray indicate abnormal findings? Choices: [A: Bilateral lung, B: Cardiac region, C: Abdomen, D: Spine]			What is the overall condition of the lungs based on the CXR? Choices: [A: Clear, B: Consolidated, C: Effused, D: Thickened]					
Ground Truth	<Answer>: Compensatory hyperinflation due to left lung removal. <Reason>: After a pneumonectomy, the remaining lung may overexpand to compensate for the loss of the left lung.			<Answer>: [A, B, C]. <Reason>: The bilateral lung shows pleural effusion, the cardiac region shows enlargement, and the abdomen shows the nasogastric tube. The spine is normal.			<Answer>: A. <Reason>: The CXR indicates that the lungs are clear of consolidation or effusion, although right pleural-based thickening at the base.					
GPT-4o-mini	<Answer>: The overexpansion of the right lung may be caused by conditions such as a right-sided pneumothorax or chronic obstructive pulmonary disease (COPD). <Reason>: Overexpansion can occur when air enters the pleural space (as in pneumothorax) or when there is an obstruction that prevents air from leaving the lung, causing it to become hyperinflated. In COPD, the lung tissue loses elasticity, leading to overinflation. <GPTScore>: 44.44			<Answer>: [A: Bilateral lung]. <Reason>: The bilateral lung region shows potential abnormalities such as opacities or other irregularities that may indicate pathology (✗). <GPTScore>: 37.50			<Answer>: B. <Reason>: The presence of increased opacity in the lung fields suggests consolidation, indicating that the lungs are not clear and may have fluid or solid material in the alveoli (✗). <GPTScore>: 33.33.					
LLaVA-Med-GEMeX	<Answer>: Possible chronic lung disease or compensatory mechanism (✓). <Reason>: Overexpansion of the lung can be due to chronic lung diseases or compensatory mechanisms, especially after a left lobectomy (✓). <GPTScore>: 66.67			<Answer>: [A, B, C]. <Reason>: The bilateral lung shows interstitial opacities (✗), the cardiac region shows enlargement (✓), and the abdomen shows the feeding tube location (✓). <GPTScore>: 87.50			<Answer>: A. <Reason>: The CXR indicates that the lungs are otherwise clear except for the noted pleural thickening (✓). <GPTScore>: 88.89					

Table 6. Challenging questions from GEMeX answered by GPT-4o-mini and our LLaVA-Med-GEMeX, **highlighting the limitations of LVLMs like GPT-4o-mini in visual reasoning, which can be enhanced by integrating knowledge from GEMeX**. Colors in outputs highlight pathological descriptions that are similar to the ground truth, though they may be either correct (✓) or incorrect (✗).

Models	Open-ended AR-Score	Closed-ended A-Score AR-Score	
		A-Score	AR-Score
LLaVA-Med-v1	73.31	56.17	62.35
LLaVA-Med-GEMeX	82.78	69.79	75.06

Table 7. Performance of LVLMs on SLAKE-CXR.

is reported in Table 7. The results indicate that our LLaVA-Med-GEMeX significantly outperforms LLaVA-Med-v1 in both tasks, demonstrating the effective X-ray knowledge acquired from GEMeX.

Case Study. In Table 6, we show some questions with outputs from both GPT-4o-mini and our fine-tuned LLaVA-Med-GEMeX for qualitative comparison. In CASE I, although GPT-4o-mini can generate a very detailed answer, it provides answers without reasoning on the visual content, resulting in a significant difference from the ground truth. In contrast, the LLaVA-Med-GEMeX offers relatively accurate visual clues and is able to provide partially correct answers (“the compensatory mechanism”), although there is a false mention of “possible chronic lung disease” when

considering the patient’s condition. In CASE II, although GPT-4o-mini can analyze images, its limited capabilities result in selecting only one option and providing a vague reason. In contrast, the LLaVA-Med-GEMeX outputs the correct options but gives an incorrect reason for one option (*i.e.*, answer “A”). In CASE III, GPT-4o-mini cannot both visually reason and output answers correctly, while the fine-tuned model can give better outputs from these two aspects. From these examples, we can conclude that some LVLMs still lack sufficient understanding of medical images. Meanwhile, while the proposed simple fine-tuning method improves performance, it is still far from fully accurate, leaving much room for further exploration. We have provided more case studies in the Appendix.

6. Conclusion

In this paper, we introduce a benchmark, GEMeX, designed to advance the field of medical VQA with two primary advantages: multimodal explainability and diverse question types. GEMeX not only provides more accessible medical explanations for patients and junior doctors but also serves as a valuable training resource for developing

next-generation medical LVLMs with enhanced instruction-following capabilities. We demonstrate the effectiveness and difficulty of the dataset through comprehensive testing of representative LVLMs as well as task-specific fine-tuning, hoping that GEMeX can promote medical VQA development and improve AI-assisted medical care.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [1](#), [2](#), [4](#), [6](#), [7](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [1](#)
- [3] Asma Alkhaldi, Raneem Alnajim, Layan Alabdullatef, Rawan Alyahya, Jun Chen, Deyao Zhu, Ahmed Alsinan, and Mohamed Elhoseiny. Minigpt-med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106*, 2024. [3](#), [6](#), [7](#), [4](#)
- [4] Seongsu Bae, Daeun Kyung, Jaehye Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#), [3](#), [4](#)
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. [4](#)
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. [6](#), [7](#)
- [7] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 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, 2021. [2](#)
- [8] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024. [7](#)
- [9] Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Le Gao, Mingpu Xu, Yue Wu, Yinwen Li, Danli Shi, and Mingguang He. Eyeqpt: Ophthalmic assistant with large language models. *arXiv preprint arXiv:2403.00840*, 2024. [3](#)
- [10] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. [3](#)
- [11] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23403–23413, 2023. [3](#)
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. [4](#)
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. [1](#)
- [14] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *Proceedings of the 2021 international conference on multimedia retrieval*, pages 456–460, 2021. [3](#)
- [15] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. [1](#)
- [16] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. [1](#), [2](#), [3](#)
- [17] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023. [2](#), [3](#), [4](#)
- [18] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lilm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. [1](#), [2](#)
- [19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoor, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. [5](#)
- [20] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las

- Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 4
- [21] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3
- [22] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multi-modal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021. 3
- [23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6
- [24] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1, 2, 3
- [25] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*, 2023. 3
- [26] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5, 6, 7, 4
- [27] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018. 1
- [28] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023. 3
- [29] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Ghollamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023. 1
- [30] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 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, 2021. 1, 2, 4, 7
- [31] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022. 3
- [32] Fenglin Liu, Shen Ge, Yuexian Zou, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*, 2022. 5
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6, 7, 4
- [34] Haotian Liu, Chunyuan Li, Qingsong Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3, 6, 7, 4
- [35] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 6, 7, 4
- [36] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022. 3
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1
- [39] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 2, 3, 6, 7, 4
- [40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambré, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4
- [41] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, pages 1–9, 2024. 3
- [42] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2, 3, 6, 7, 4
- [43] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024. 4
- [44] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imangenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. 2, 3

- [45] Li Xu, Bo Liu, Ameer Hamza Khan, Lu Fan, and Xiao-Ming Wu. Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark. *arXiv preprint arXiv:2306.06494*, 2023. 3
- [46] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. LvLM-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 1
- [47] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 6, 7, 4
- [48] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 3
- [49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6
- [50] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1, 2, 3
- [51] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024. 2, 3
- [52] Xiangyu Zhao, Bo Liu, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. Easygen: Easing multimodal generation with bidiffuser and llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1351–1370, 2024. 1
- [53] Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. Chatcad+: Towards a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, 2024. 2
- [54] Jue Xiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiyi Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2304.10691*, 2023. 3
- [55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 6, 7, 4
- [56] Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. Medrg: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*, 2024. 3

GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis

Supplementary Material

	Chest Imagenome	Ours
Reserve	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc
Incorporate	left upper lung zone left apical zone	left upper lung zone
	right upper lung zone right apical zone	right upper lung zone
	mediastinum upper mediastinum	mediastinum
	right lower lung zone right costophrenic angle	right lower lung zone
	left lower lung zone left costophrenic angle	left lower lung zone
	cardiac silhouette cavoatrial junction right atrium	cardiac silhouette
Delete	carina right clavicle left clavicle aortic arch	-
Merge	left lung + right lung	bilateral lung
	left upper + right upper	bilateral upper lung zone
	left mid + right mid	bilateral mid lung zone
	left lower + right lower	bilateral lower lung zone
	left hilar + right hilar	bilateral hilar structures
	left hemidiaphragm + right hemidiaphragm	bilateral hemidiaphragm
	left mid + left lower	left mid-to-lower lung zone
	right mid + right lower	right mid-to-lower lung zone
	left mid + left upper	left mid-to-upper lung zone
	right mid + right upper	right mid-to-upper lung zone
	left mid-to-lower + right mid-to-lower	bilateral mid-to-lower lung zone
	left mid-to-upper + right mid-to-upper	bilateral mid-to-upper lung zone
Sum	29	30

Table 8. Anatomical regions transformation from the Chest Im-aGenome to our GEMeX version. The left column indicates the detailed operation.

7. GEMeX Construction Details

7.1. Transformation and Distribution of Anatomical Regions

As we said in Section 3.1, we provide detailed operations to transform anatomical regions from Chest ImaGenome to our GEMeX. The process is summarized in Table 8. The resulting anatomical region distribution corresponding to each sentence is shown in Figure 5. Overall, there are 30 regions, and the merged area occupies a large proportion, such as “bilateral lung” and “bilateral hilar structures”.

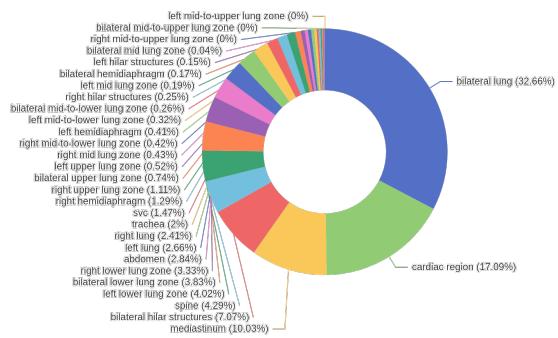


Figure 5. Distribution of anatomical regions corresponding to each sentence after transformation from the Chest Imagenet dataset.

7.2. Prompt for Re-grounding Report

Here, we provide detailed instructions to re-ground reports with medical LLM, as we elaborated in Section 3.1. The prompt is shown in Table 9, where we add clinical guidance (like (2) (4) (5) (6)) and split and re-written requirements (*e.g.*, (3)) to ensure correct sentence-region correspondence. Moreover, we provide some manually labeled pairs as demonstrations (*i.e.*, “here are some cases”) for in-context learning, aiming to improve overall performance.

7.3. Prompt for VQA Generation

In Table 10, we provide a detailed prompt to guide GPT-4o in VQA generation. Specifically, for each CXR, we generate 3 open-ended questions, 2 closed-ended questions, 3 single-choice questions, and 3 multi-choice questions, containing diverse content, like “abnormality”, “disease”, etc. GPT-4o is required to provide detailed reasoning and specify relevant visual regions (rule (2)). Additionally,

messages = [“role”：“**system**”, “content”：
f““You are a helpful chest X-ray radiologist. Given an input sentence, your task is to map it to an anatomical region on X-ray for better observation from a predefined list [right lung, cardiac silhouette,..., bilateral lower lung zone].
Here are some rules:
(1) If there is no corresponding region for this sentence, leave it out.
(2) If the sentence describes the overall anatomical characteristics without specifying a particular region, you can choose “bilateral lung” as its region. For example, “No focal consolidation, pleural effusion or pneumothorax is present.”:“bilateral lung”.
(3) One sentence can only correspond to one region. If a sentence’s main symptom involves several anatomical regions, rephrase it into multiple sentences with corresponding regions. Note that all derived sentences must be syntactically complete, not phrases (i.e., containing a subject and a predicate at least). For example: “The cardiomedastinal silhouette is normal.” can be segmented into “The cardiac silhouette is normal.”:“cardiac region”, “The mediastinal silhouette is normal.”:“mediastinum”, where “cardiomedastinal” corresponds to the “cardiac region” and “mediastinum”.
(4) Small (tiny) pleural effusion (fluid) usually appears in the lower lung zone, a moderate pleural effusion appears in the mid-to-lower lung zone, and a large (substantial) pleural effusion can even occupy the entire lung. If the severity (like small, moderate and large) is not indicated, output the left lung or right lung directly.
(5) The main anatomical region for observing pulmonary venous is the bilateral hilar structures on the X-ray.
(6) The region where the atrium and ventricle can be observed is the cardiac region.
Here are some cases: (1)... (2)... (3)... (4)...
Organize your output in a json formatted as DictStr(sentence):Str(region), without other words.””]

messages += [“role”：“**user**”, “content”：
“Input: “Bibasilar atelectasis is seen without discrete focal consolidation.””]

Table 9. Our proposed prompt guided by radiologist feedback for refining sentence-region pairs.

messages = [“role”：“**system**”, “content”：
f““You are a chest X-ray AI assistant, and you are seeing a frontal view chest X-ray image, described by several phrases with visual regions. Generate **3** open-ended questions, **2** closed-ended questions, **3** single-choice questions, and **3** multi-choice questions about this chest X-ray. Format your output in JSON format.
Here are some rules:
(1) Include questions asking about the visual content of the image, containing abnormality, disease, location, severity, cause of disease, size, and implication. For a CXR, the types of questions generated need to be diverse. Do not ask any questions that cannot be answered confidently.
(2) For each question, generate its type (abnormality, location, ...), provide the answer, explain the reason for obtaining such answer, and output the corresponding visual regions as a visual clue.
(3) For open-ended questions, the answers must be concise. You should generate detailed reasons based on the provided CXR phrases and your medical knowledge. Do not refer to the text description in your questions or answers.
(4) Avoid questions that cannot be answered by looking at the given CXR image itself, such as asking about changes/comparisons from previous scans, asking about staff notifications, or asking about view types or other scans.
Here is one example:
Chest X-ray: {...}, One open-ended question can be: {...}, One closed-ended question can be: {...}, One single-choice question can be: {...}, One multi-choice question can be: {...}

messages += [“role”：“**user**”, “content”：
“Chest X-ray: There is also fullness of the right hilum which is new. [visual location: right hilar structures] ...”

Table 10. Our designed prompt for generating groundable and explainable medical VQA, using a grounded report as input.

comparison-based questions are excluded, as only a single CXR is provided (rule (3)). To better align with our objectives, we incorporate manually crafted questions as demonstrations (the inputs after “Here is one example”).

8. Benchmark Details

8.1. Distribution Analysis of Question and Reason Lengths Across Data Splits

We provide a detailed distribution of question and reason lengths across data splits, as presented in Figure 6. It can

be seen that under different data splits, the distributions of question (reason) lengths are generally similar. Furthermore, from the perspective of reason, the reason lengths mostly fall between 60 and 150, demonstrating the level of detail in the reasons as textual explanations.

8.2. Word Frequency Analysis of Questions and Reasons

Besides providing length distribution, we also explore the frequency of words from both questions and reasons, as shown in Figure 7. From the left part (regarding questions),

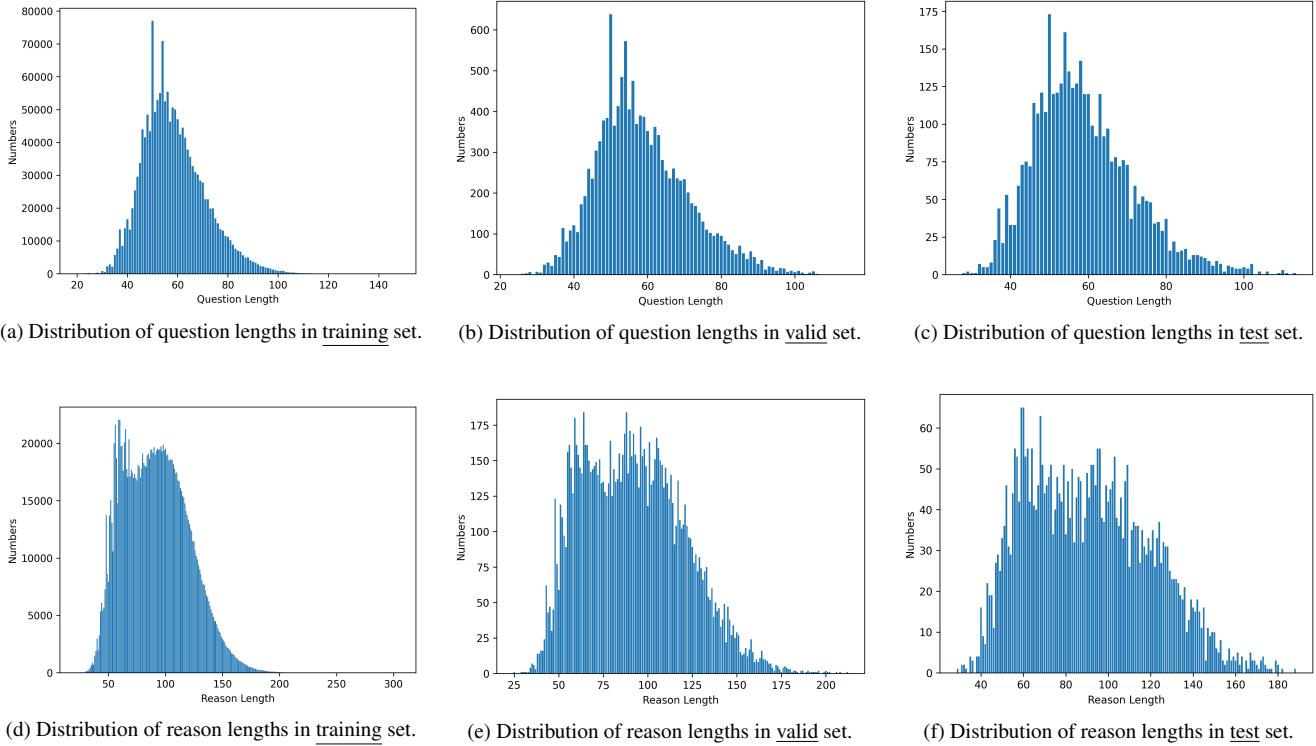


Figure 6. Detailed distribution of question lengths and reason lengths across data splits.



Figure 7. Visualization of word frequencies in questions (left) and reasons (right).

we can observe that the majority of words are question-related terms, such as “what”, “which”, “is”, and “are”. Additionally, some disease-related terms, such as “abnormality”, “findings”, and “pleural effusion”, are also quite common. Lastly, content words related to the questions, such as “regions”, “evidence”, and “size”, are frequently mentioned. These demonstrate the diversity of questions; On the right, we show the word cloud of reasons. It can be seen that the vocabulary mainly falls into two categories: one is related to diseases or anatomical regions, such as “normal”, “heart”, and “pleural”, and the other consists of words used to convey explanations, such as “indicates” and “states”.

8.3. Input Samples for the Proposed Fine-tuning

We here provide some input samples for a better understanding of how to fine-tune LLaVA-Med [26] on our GMeX. An input sample of open-ended questions can be seen in Table 11; An input sample of closed-ended questions can be seen in Table 12; An input sample of single-choice questions can be seen in Table 13; An input sample of multi-choice questions can be seen in Table 14.

8.4. Fine-tuning Details

We fine-tune both the visual projection layers and the LLM components of LLaVA-Med-v1 (after stage II) by calcu-

lating the auto-regressive loss to predict the assistant’s responses and the dialogue termination token <STOP>. Particularly, the model is trained for 3 epochs on four NVIDIA H100 GPUs with a batch size of 64, taking around 54 hours. The network is warmed up in the first 0.03 epochs with a linear learning rate from 3e-7 to 2e-5, which further decays by cosine schedule. The optimizer is AdamW. To accelerate training, we employ the Fully Sharded Data Parallel (FSDP) mechanism, the bf16 (Brain Floating Point) data format, and gradient checkpointing.

8.5. LVLMs Introduction

Besides fine-tuning a task-oriented model, we perform a zero-shot evaluation on our GEMeX dataset across the other 12 LVLMs, with 7 in the general domain and the other 5 in the medical domain:

- **In the General Domain:** LLaVA-v1 [34] and Mini-GPT4-v1 [55] are two pioneering works, achieving remarkable results in multimodal tasks; mPLUG-Owl [47] is a multimodal model from the mPLUG series focused on visual-language tasks; LLaVA-v1.5 [33] is an improved version of LLaVA-v1 with enhancements in multimodal alignment, accuracy, and efficiency; Deepseek-VL [35] uses the SigLIP-L and SAM-B as the hybrid vision encoder and the DeepSeek-LLM as the base LLM for solving multimodal tasks; Qwen-VL-Chat [5] originates from the Qwen family, possessing capabilities such as multilingual dialogue and multi-image interleaved dialogue; GPT-4o-mini is a smaller, optimized version of GPT-4 [1] intended for lighter computational environments. Note that we did not test GPT-4o because its safety protection policy prohibits it from analyzing medical images.
- **In the Medical Domain:** LLaVA-Med-v1 [26] is designed specifically for medical applications based on LLaVA-v1 and LLaVA-Med-v1.5 [26] is an advanced version; MiniGPT-Med [3] is an medical version of Mini-GPT4 [55]; XrayGPT [39] is a specialized GPT model for interpreting chest X-rays; RadFM [42] is a radiology foundation model.

To make a fair comparison, the evaluated models (except GPT-4o-mini and RadFM (with MedLLaMA-13B [43])) are based on 7B-LLMs in this section. Specifically, LLaVA-v1, LLaVA-Med-v1, and Mini-GPT4-v1 are based on Vicuna-v0-7B [12] while LLaVA-v1.5 and XrayGPT are based on Vicuna-v1-7B; LLaVA-Med-v1.5 is built upon Mistral-7B-Instruct-v0.2 [20]; mPLUG-Owl is using LLaMA-7B [40]; Deepseek-VL [35] is based on DeepSeek-LLM-7B; and Qwen-VL-Chat [5] is based on Qwen-7B. All models’ configurations are set according to their open-source codes.

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions. Input an open-ended question, and the assistant will output its answer with a detailed reason and corresponding visual location.\n\n### Human:<image>\n What abnormalities are observed in the left lower lung zone?\n### Assistant: <answer>Linear atelectasis. <reason>The CXR indicates that the lungs are clear except for linear atelectasis located specifically at the left base. <location>[[126, 110, 203, 167]]\n###

Table 11. An input sample of open-ended questions

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions. Input a closed-ended question, and the assistant will output its answer (yes or no) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n Are there any lung abnormalities present in this CXR?\n### Assistant: <answer>No. <reason>The CXR clearly shows that the lungs are clear, indicating no lung abnormalities. <location>[[30, 34, 185, 178]]\n###

Table 12. An input sample of closed-ended questions

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions. Input a single-choice question, and the assistant will output its answer (an option) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n Which of the following is absent in this CXR? <choices>: [A: Pulmonary edema, B: Pleural effusion, C: Pneumothorax, D: All of the above]\n### Assistant: <answer>D <reason>The CXR shows that there is no pulmonary edema, effusion, or pneumothorax present. <location>[[48, 48, 175, 180]]\n###

Table 13. An input sample of single-choice questions

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions. Input a multi-choice question, and the assistant will output its answer (some options) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n What abnormalities are mentioned regarding the lung fields? <choices>: [A: Clear lung fields, B: Atelectasis, C: Effusions, D: Congestion]\n### Assistant: <answer>[B, C] <reason>The CXR shows bibasilar atelectasis and small pleural effusions. There is no impressive congestion shown. <location>[[26, 119, 217, 183]]\n###

Table 14. An input sample of multi-choice questions

9. More Case Studies

Here, we present more questions with answers from GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX, for better understand our dataset GEMeX and the corresponding performance of LVLMs.

- We first present some cases of **open-ended questions**, as in Table 15. It can be seen that our fine-tuned model can generally provide correct (or partially correct) answers and identify relatively accurate visual locations. However, other models fail to deliver both precise textual answers and accurate visual positions simultaneously.
- Furthermore, we provide some cases from **closed-ended questions** in Table 16. Although LLaVA-Med can correctly answer the first two questions, it fails to provide visual grounding. For the third question, GPT-4o-mini provides a correct answer, but there is a discrepancy between its grounded visual location (mediastinum) and the ground truth (cardiac region). In contrast, our fine-tuned model can provide both correct answers and accurate visual grounding.
- Next, we show three cases of **single-choice questions**, presented in Table 17. Overall, GPT-4o-mini and LLaVA-Med demonstrate insufficient image understanding capabilities. For instance, in the third example (CASE III), both models incorrectly identify pleural effusion on both sides, whereas it is actually present only in the left lower lung. In comparison, the fine-tuned model shows significant improvement in visual understanding, as evidenced by the grounding results.
- Finally, some cases from **multi-choice questions** are illustrated in Table 18. It can be observed that multi-choice questions are generally more challenging. GPT-4o-mini can only partially identify the correct options; LLaVA-Med tends to answer questions directly based on the question. For example, in all CASEs, it outputs all textually corresponding answers and provides seemingly reasonable explanations, but some of these answers are incorrect when judged based on visual content; The fine-tuned model may also make analytical errors. In CASE III, it incorrectly determines the presence of pleural effusion.

Open-ended questions from GEMeX:

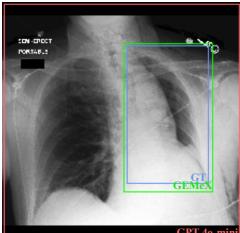
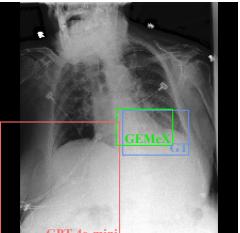
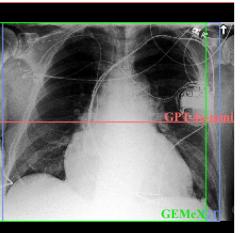
 CASE I  CASE II  CASE III			
	(Bbox predicted by each model)	(Bbox predicted by each model)	(Bbox predicted by each model)
Question	What might be the cause of the over-expansion of the right lung?	What disease is indicated at the left lung base?	What is the condition of the lungs in this CXR?
Ground Truth	<Answer>: Compensatory hyperinflation due to left lung removal. <Reason>: After a pneumonectomy, the remaining lung may overexpand to compensate for the loss of the left lung.	<Answer>: Atelectasis and/or pulmonary edema. <Reason>: The increased opacity at the left lung base suggests either atelectasis or a combination of atelectasis and pulmonary edema.	<Answer>: Clear. <Reason>: The CXR indicates that the lungs appear clear without definite signs of pneumonia or congestive heart failure (CHF).
GPT-4o-mini	<Answer>: The overexpansion of the right lung may be caused by conditions such as a right-sided pneumothorax or chronic obstructive pulmonary disease (COPD) (✗). <Reason>: Overexpansion can occur when air enters the pleural space (as in pneumothorax) or when there is an obstruction that prevents air from leaving the lung, causing it to become hyperinflated. In COPD, the lung tissue loses elasticity, leading to overinflation. <GPTScore>: 44.44	<Answer>: Possible pneumonia or pleural effusion (✗). <Reason>: The left lung base shows increased opacity, which may indicate fluid accumulation (pleural effusion) or consolidation (pneumonia). These conditions often present as areas of increased density on chest X-rays. <GPTScore>: 55.56	<Answer>: The lungs appear to have a significant opacity, suggesting possible fluid accumulation or consolidation (✗). <Reason>: The central area of the chest X-ray shows increased density, which may indicate the presence of pleural effusion or other pathological processes affecting lung visibility. The clear outlines of the lungs are obscured, which is not typical in a healthy CXR. <GPTScore>: 33.33
LLaVA-Med	<Answer>: The overexpansion of the right lung with a mediastinal shift to the left could be caused by various factors, such as a buildup of air, fluid, or pressure within the lung cavity. This can be due to conditions like pneumothorax (collapsed lung), pleural effusion (fluid buildup around the lungs), or a mass causing pressure on the lung (✗). It's important to consider the patient's clinical history, symptoms, and additional imaging findings to determine the exact cause of the overexpansion and mediastinal shift. <GPTScore>: 55.56	<Answer>: The image shows left lung base fibrocystic disease (✗), which is a lung condition characterized by the presence of fibrous and cystic tissue in the lung base. <GPTScore>: 22.22	<Answer>: In this post-operative chest X-ray (CXR), the lungs appear to be clear and without any signs of infection (✓). <GPTScore>: 77.78
LLaVA-Med-GEMeX	<Answer>: Possible chronic lung disease or compensatory mechanism (✓). <Reason>: Overexpansion of the lung can be due to chronic lung diseases or compensatory mechanisms, especially after a left lobectomy (✓). <GPTScore>: 66.67	<Answer>: Atelectasis (✓). <Reason>: The CXR shows moderate atelectasis at the left lung base (✓), which is a collapse or incomplete expansion of the lung. <GPTScore>: 77.78	<Answer>: Clear. <Reason>: The CXR indicates that the lungs are clear with normal pulmonary vasculature (✓). <GPTScore>: 88.89

Table 15. Open-ended questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

Closed-ended questions from GEMeX:

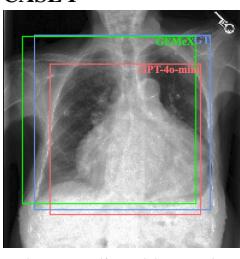
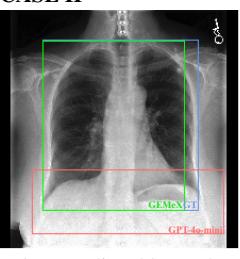
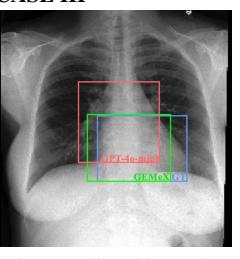
	CASE I 	CASE II 	CASE III 
	(Bbox predicted by each model)	(Bbox predicted by each model)	(Bbox predicted by each model)
<i>Question</i>	Are there signs of pneumonia in this CXR?	Is there any pleural effusion present in the CXR?	Is the heart size abnormal in this CXR?
<i>Ground Truth</i>	<Answer>: No. <Reason>: The CXR explicitly states that there is no evidence of pneumonia, indicating clear lung fields.	<Answer>: No. <Reason>: The CXR explicitly states that there is no pleural effusion.	<Answer>: No. <Reason>: The CXR indicates that the heart size is within normal limits.
GPT-4o-mini	<Answer>: Yes. <Reason>: The chest X-ray shows areas of increased opacity, which may indicate the presence of fluid or consolidation (✗) typically associated with pneumonia. <GPTScore>: 37.50	<Answer>: Yes. <Reason>: The presence of pleural effusion is indicated by the blunting of the costophrenic angles and the increased opacity in the lower lung fields (✗) , suggesting fluid accumulation in the pleural space. <GPTScore>: 20.00	<Answer>: No. <Reason>: The heart size appears to be within normal limits (✓) based on the overall silhouette and the ratio of the heart to the thoracic cavity. <GPTScore>: 100.00
LLaVA-Med	<Answer>: No, there are no signs of pneumonia in this chest X-ray (✓) . <GPTScore>: 87.50	<Answer>: No, there is no pleural effusion present in the chest X-ray (CXR) (✓) . <GPTScore>: 100.00	<Answer>: Yes, the heart size in this chest X-ray appears to be abnormally enlarged (✗) . This is referred to as cardiomegaly. <GPTScore>: 20.00
LLaVA-Med-GEMeX	<Answer>: No. <Reason>: The CXR indicates no focal consolidation, which rules out pneumonia (✓) . <GPTScore>: 100	<Answer>: No. <Reason>: The CXR clearly states that there is no pleural effusion (✓) . <GPTScore>: 100.00	<Answer>: No. <Reason>: The heart is described as normal in size (✓) . <GPTScore>: 100.00

Table 16. Closed-ended questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

Single-choice questions from GEMeX:

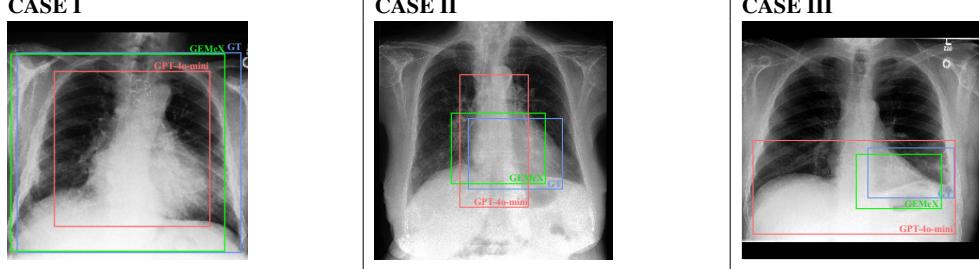
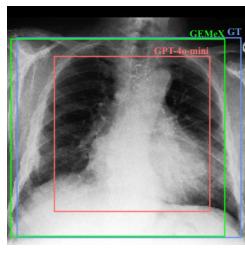
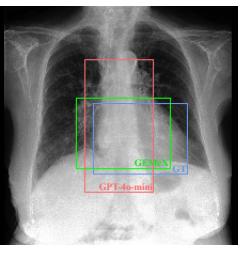
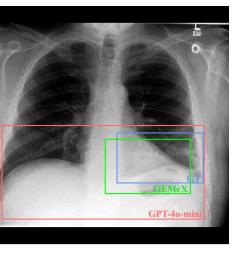
			
	CASE I  (Bbox predicted by each model)	CASE II  (Bbox predicted by each model)	CASE III  (Bbox predicted by each model)
Question	What is the overall condition of the lungs based on the CXR? Choices: [A: Clear, B: Consolidated, C: Effused, D: Thickened]	What does the CXR indicate about the heart size? Choices: [A: Normal, B: Mild cardiomegaly, C: Moderate cardiomegaly, D: Severe cardiomegaly]	Which regions are assessed for pleural effusion in this CXR? Choices: [A: Right lower lung zone, B: Left lower lung zone, C: Cardiac region, D: Bilateral lung]
Ground Truth	<Answer>: A. <Reason>: The CXR indicates that the lungs are clear of consolidation or effusion, although right pleural-based thickening at the base.	<Answer>: C. <Reason>: The CXR indicates unchanged moderate cardiomegaly.	<Answer>: B. <Reason>: The pleural effusion is specifically noted in the left lower lung zone.
GPT-4o-mini	<Answer>: B. <Reason>: The presence of increased opacity in the lung fields suggests consolidation, indicating that the lungs are not clear and may have fluid or solid material in the alveoli (✗). <GPTScore>: 33.33	<Answer>: B. <Reason>: The CXR shows an enlarged heart silhouette, indicating mild cardiomegaly. The heart appears larger than normal but not excessively so (✗). <GPTScore>: 55.56	<Answer>: D. <Reason>: Pleural effusion can accumulate in the lower lung zones, and both sides of the chest. The bilateral lung regions are crucial for evaluating pleural effusion (✗). <GPTScore>: 66.67
LLaVA-Med	<Answer>: The chest X-ray (CXR) shows bilateral diffuse interstitial thickening, which means that both lungs have a thickened appearance in their interstitial areas (✗). This finding can be associated with various lung conditions, such as infections, inflammation, or fibrosis. It is important to consider the patient's clinical history and symptoms, as well as consult a healthcare professional for a thorough evaluation and proper diagnosis of the underlying cause of these findings. <GPTScore>: 37.50	<Answer>: The CXR indicates that there is mild cardiomegaly, which means that the heart appears slightly enlarged (✗). <GPTScore>: 44.44	<Answer>: In this chest X-ray (CXR), the regions assessed for pleural effusion include the right lower lung zone, the left lower lung zone, the cardiac region, and the bilateral lung (✗). <GPTScore>: 50.00
LLaVA-Med-GEMeX	<Answer>: A. <Reason>: The CXR indicates that the lungs are otherwise clear except for the noted pleural thickening (✓). <GPTScore>: 88.89	<Answer>: C. <Reason>: The CXR indicates moderate cardiomegaly (✓). <GPTScore>: 100.00	<Answer>: B. <Reason>: The CXR specifically indicates that trace pleural effusion is difficult to exclude in the left lower lung zone (✓). <GPTScore>: 87.50

Table 17. Single-choice questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

Multi-choice questions from GEMeX:

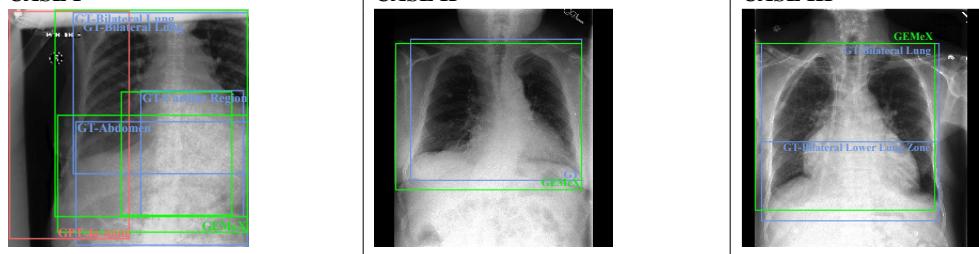
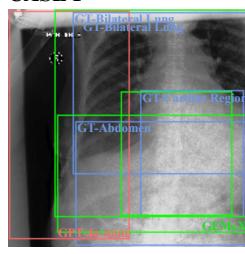
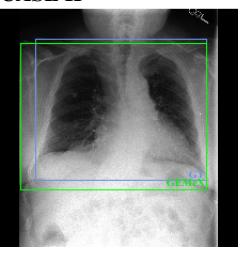
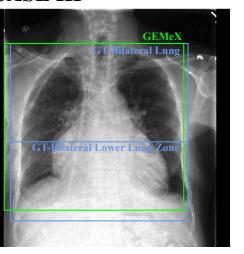
			
	CASE I  (Bbox predicted by each model)	CASE II  (Bbox predicted by each model)	CASE III  (Bbox predicted by each model)
Question	What regions on the X-ray indicate abnormal findings? Choices: [A: Bilateral lung, B: Cardiac region, C: Abdomen, D: Spine]	Which findings indicate abnormal lung conditions? Choices: [A: Mild interstitial abnormality, B: Hyperinflation, C: Focal consolidation, D: No pleural effusion]	Which observations rule out serious lung conditions in this CXR? [A: No pneumonia, B: No lung nodules, C: No pleural effusions, D: No lymphadenopathy]
Ground Truth	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows pleural effusion , the cardiac region shows enlargement, and the abdomen shows the nasogastric tube . The spine is normal.	<Answer>: [A, B]. <Reason>: Mild interstitial abnormality and hyperinflation indicate abnormal lung conditions . There is no focal consolidation.	<Answer>: [A, B, D] <Reason>: The CXR shows no pneumonia, lung nodules, or lymphadenopathy , which rules out serious lung conditions. However, there are mild bilateral pleural effusions .
GPT-4o-mini	<Answer>: [A: Bilateral lung]. <Reason>: The bilateral lung region shows potential abnormalities such as opacities or other irregularities that may indicate pathology (✗) . <GPTScore>: 37.50	<Answer>: [A, B, C]. <Reason>: Mild interstitial abnormality, hyperinflation (✓) , and focal consolidation (✗) are all findings that indicate abnormal lung conditions. <GPTScore>: 66.67	<Answer>: [A, B, C, D]. <Reason>: The absence of pneumonia, lung nodules, pleural effusions (✗), and lymphadenopathy (✓) in the chest X-ray suggests that there are no serious lung conditions present. Each of these findings is associated with significant lung pathology, and their absence helps rule out serious issues. <GPTScore>: 62.50
LLaVA-Med	<Answer>: The chest X-ray shows abnormal findings in multiple regions. The left side of the image indicates bilateral lung abnormalities , which could be related to lung issues such as infection, inflammation, or other lung conditions (✗). The right side of the image shows abnormal findings in the cardiac region, which may suggest issues with the heart or surrounding structures (✓). The lower part of the image reveals abnormal findings in the abdomen , which could be related to gastrointestinal or other abdominal issues (✓). Lastly, the spine region also shows abnormal findings, which could be related to spinal issues or other back-related problems. <GPTScore>: 50.00	<Answer>: The chest x-ray shows findings A, B, and C, which indicate abnormal lung conditions. These findings are likely a combination of interstitial thickening, hyperinflation (✓), and focal consolidation (✗) . These abnormalities can be associated with various lung conditions, such as chronic obstructive pulmonary disease (COPD), asthma, or other respiratory disorders. <GPTScore>: 55.56	<Answer>: The image shows a chest X-ray that demonstrates the absence of certain conditions, which are marked as B on the image. These conditions include pneumonia, lung nodules, lymphadenopathy (✓), and pleural effusions (✗) . This means that the chest X-ray does not show any signs of these lung conditions, ruling them out based on the image. <GPTScore>: 62.50
LLaVA-Med-GEMeX	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows interstitial opacities (✗) , the cardiac region shows enlargement (✓) , and the abdomen shows the feeding tube location (✓) . <GPTScore>: 87.50	<Answer>: [A, B]. <Reason>: The CXR shows mild interstitial abnormality and hyperinflation (✓) as abnormal lung conditions. <GPTScore>: 88.89	<Answer>: [A, B, C, D]. <Reason>: The CXR indicates the absence of pneumonia, lung nodules, pleural effusions (✗), and lymphadenopathy (✓) . <GPTScore>: 62.50

Table 18. Multi-choice questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.