Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale

Junying Chen[1,2], **Chi Gui**[2], **Ruyi Ouyang**[2], **Anningzhe Gao**[1,2], **Shunian Chen**[1,2]
Guiming Hardy Chen[1,2], **Xidong Wang**[1,2], **Zhenyang Cai**[1,2], **Ke Ji**[1,2]
Guangjun Yu[1,2,3], **Xiang Wan**[1,2,3], **Benyou Wang**[1,2*]
[1] Shenzhen Research Institute of Big Data, [2] The Chinese University of Hong Kong, Shenzhen
[3] National Health Data Institute, Shenzhen
wangbenyou@cuhk.edu.cn

Abstract

The rapid development of multimodal large language models (MLLMs), such as GPT-4V, has led to significant advancements. However, these models still face challenges in medical multimodal capabilities due to limitations in the quantity and quality of medical vision-text data, stemming from data privacy concerns and high annotation costs. While pioneering approaches utilize PubMed's large-scale, de-identified medical image-text pairs to address these limitations, they still fall short due to inherent data noise. To tackle this, we refined medical image-text pairs from PubMed and employed MLLMs (GPT-4V) in an *'unblinded'* capacity to denoise and reformat the data, resulting in the creation of the **PubMedVision** dataset with 1.3 million medical VQA samples. Our validation demonstrates that: **(1)** PubMedVision can significantly enhance the medical multimodal capabilities of current MLLMs, showing significant improvement in benchmarks including the MMMU Health & Medicine track; **(2)** manual checks by medical experts and empirical results validate the superior data quality of our dataset compared to other data construction methods. Using PubMedVision, we train a 34B medical MLLM **HuatuoGPT-Vision**, which shows superior performance in medical multimodal scenarios among open-source MLLMs. Our code and data are available at https://github.com/FreedomIntelligence/HuatuoGPT-Vision.

1 Introduction

Multimodal Large Language Models (MLLMs), such as GPT4-V, show limited performance in medical applications, particularly in lacking visual knowledge specific to the medical domain (Yan et al., 2023; Jin et al., 2024). Although there are some small-scale, high-quality datasets containing medical visual knowledge (Lau et al., 2018; Liu et al., 2021; He et al., 2020), scaling them up is challenging. Additionally, there are privacy and licensing issues associated with medical data, further complicating matters.

Pioneering works (Zhang et al., 2023c; Li et al., 2023a; Wu et al., 2023) utilize PubMed[1] for larger-scale training for medical vision-language alignment. PubMed is favored because it contains medical images and surrounding text, which (i) encapsulate the forefront of human wisdom in medicine and (ii) are well-de-identified (Lin et al., 2023). However, models trained on PubMed are unsatisfactory, as they perform poorly compared to general MLLMs on medical multimodal tasks (Hu et al., 2024; Xia et al., 2024). This can be attributed to data noise in PubMed, which significantly affects multimodal performance (Liu et al., 2024; Chen et al., 2024).

Concurrently, LLaVA-Med (Li et al., 2023a) uses a "*blind*" Large Language Model (LLM) to generate Visual Question Answering (VQA) from the contextual text of PubMed images, achieving notable results. However, this approach might overlook visual information inherent in the medical images themselves as LLMs cannot perceive images as input, probably leading to the generation of misinterpreted or irrelevant answers. Moreover, LLaVA-Med is limited to 56K medical VQA entries. Thus, creating a higher-quality and larger-scale vision-language alignment dataset for medicine is essential.

To close this gap, we meticulously select high-quality medical image-text pair from PubMed, employing a proposed refined pipeline. Utilizing 914,960 refined medical images and their corresponding text, we apply GPT-4V as the "*unblinded*" reformatter, contrasting the "*blinded*" reformatting used in previous works (Li et al., 2023a; Wu et al.,

*Benyou is the corresponding author.

[1]PubMed is a free search engine that primarily accesses the MEDLINE database, containing references and scientific papers on life sciences and biomedical topics.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7346–7370
November 12-16, 2024 ©2024 Association for Computational Linguistics

2023; Zhang et al., 2023c), to denoise the PubMed data. Our method generates more aligned medical VQA data for medical multimodal alignment. Consequently, we constructed a high-quality multimodal medical dataset with *1.3 million* samples and name it as **PubMedVision**.

Our experiments validated **PubMedVision** in two key aspects: (1) It significantly enhances the medical multimodal capabilities of MLLMs, showing notable improvement in benchmarks such as MMMU Health & Medicine. LLaVA-v1.5-LLaMA-3-8B achieves the strongest performance among open-source MLLMs with PubMedVision ; (2) Manual checks by medical experts and empirical results confirmed the superior data quality of PubMedVision compared to current data construction methods. The contributions of this paper are summarized as follows:

1. **Unblinded Data Reformatting for Medical Multimodality.** We propose leveraging "*unblinded*" MLLMs to reformat PubMed image-text pairs to construct a better-aligned medical VQA dataset. Expert reviews and empirical tests show that this method yields higher-quality data, improving MLLM training.

2. **PubMedVision: A Large-scale, High-quality Medical Multimodal Dataset.** With the MLLM-powered reformatted method, we bulid PubMedVision, containing 1.3 million medical VQA entries for visual alignment. Experiments demonstrate that PubMedVision significantly enhances MLLMs' medical multimodal capabilities, enabling models like *LLaVA-1.5-LLaMA-3-8B* to outperform other general and medical open-source MLLMs.

3. **HuatuoGPT-Vision: A Medical MLLM.** Using PubMedVision, we trained HuatuoGPT-Vision, a 34B parameter medical MLLM. HuatuoGPT-Vision demonstrate superior performance on multiple medical multimodal benchmarks among open-source models.

2 Medical Visual Alignment in MLLMs

2.1 Existing Medical VQA Data

Table 1 compares existing medical VQA datasets, which are crucial for image-text alignment and instruction following in medical MLLMs. Early datasets like VQA-RAD, SLAKE, and Path-VQA are limited by their small size (less than 20K entries) and their exclusive focus on radiology. PMC-CaseReport (Wu et al., 2023), PMC-VQA (Zhang et al., 2023c), and LLaVA-Med leverage PubMed medical images to scale data and employ LLMs to reformat contextual text into VQA. However, these datasets also suffer from limited quantity and are prone to misinterpretation and misalignment due to the 'blinded' nature of the LLMs. In contrast, we aim to construct a larger-scale, high-quality medical VQA dataset, PubMedVision.

2.2 Medical Visual Alignment through the Lens of Data Engineering

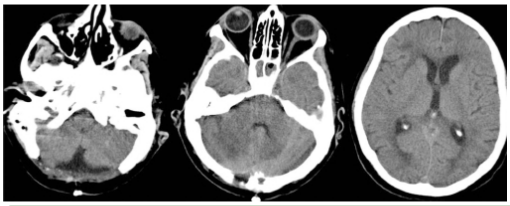
Visual Knowledge Alignment Current MLLMs typically adapt a text-only LLM with a visual encoder (Liu et al., 2024; Li et al., 2023b). Therefore, *alignment* involves injecting image knowledge into LLMs, aligning images with the language understanding of LLMs. This paper explores the injection of extensive medical visual knowledge from PubMed into MLLMs, as PubMed is a leading repository of advanced medical research with well-de-identified medical images.

Data Noises in PubMed Although existing work (Wu et al., 2023; Li et al., 2023a; Zhang et al., 2023c) utilize PubMed, it has not been entirely satisfactory, as they still lag behind many general-purpose MLLMs in medical vision (Hu et al., 2024; Xia et al., 2024). We attribute it to the data noises in PubMed. The text surrounding the image in PubMed papers does not always well-describe the image. While relevant, this text does not necessarily facilitate effective visual alignment.

The Efforts to Improve Data Quality Sourced from PubMed The original data is not always suitable for training, as seen in reformatting alignment (Fan et al., 2024). Compared to Native Captions in PubMed, existing work uses text-only LLMs to reformat these captions of images, denoted as LLM-Reformatted . This can result in misinterpreted or misaligned text for the images due to the blined LLM. To solve this, we propose using a *multimodal* LLM, called MLLM-Reformatted . Additionally, we compare with GPT4v-Distill , a popular approach to distill GPT-4V in general multimodal fields, such as ShareGPT4V (Chen et al., 2023b) and ALLaVA-4V (Chen et al., 2024). For GPT4v-Distilled, we provide only images to GPT-4V to generate a medical description.

	Data Size	Modality	Synthetic?	Uses LLMs?	Uses MLLMs?	Source
VQA-RAD (Lau et al., 2018)	3,515	Radiology	✓	×	×	Websites
SLAKE (Liu et al., 2021)	14,028	Radiology	✓	×	×	Websites
PathVQA (He et al., 2020)	17,325	Pathology	✓	×	×	Books
PMC-CaseReport (Wu et al., 2023)	54,341	Radiology	✓	✓	×	PubMed
PMC-VQA (Zhang et al., 2023c)	176,919	Multimodal	✓	✓	×	PubMed
LLaVA-Med VQA (Li et al., 2023a)	56,702	Multimodal	✓	✓	×	PubMed
PubMedVision (Ours)	1,294,062	Multimodal	✓	×	✓	PubMed

Table 1: Comparison of Medical VQA Datasets



Native-Caption ☹

Postoperative cranial non contrast-enhanced CT-scan demonstrating: 1) a complete resorption of the cerebellar haemorrhage; 2) a regression of posterior fossa swelling; 3) a visible fourth ventricle associated with a regression of the hydrocephalus. Control CT-scan ~~, performed one week later~~ , demonstrated a complete resorption of the haemorrhage and a regression of posterior fossa oedema and hydrocephalus~~ (Fig. 2)~~.

LLM-Reformatted ☹

The image is ~~a postoperative cranial non-contrast-enhanced CT scan slide~~, ~~displaying significant improvements following a medical intervention.~~ Firstly, it shows that the cerebellar hemorrhage has been fully resorbed, indicating no residual bleeding in the area. Secondly, there is a noticeable reduction in the swelling of the posterior fossa. Lastly, the fourth ventricle is clearly visible, which correlates with the observed reduction in hydrocephalus. This image illustrates successful outcomes of the treatment procedures in resolving these critical conditions.

GPT4v-Distill ☹

The medical image provided is a computed tomography (CT) scan of the head, showcasing three axial slices at different levels:
1. Left Image (~~Axial CT Scan with Contrast~~). This slice captures the basal cisterns and surrounding structures. ~~There is enhancement observed in the blood vessels due to contrast administration.~~ The detailed view of bony structures such as the sphenoid sinus and temporal bones can be seen. 2. Middle Image (~~Axial CT Scan with Contrast~~). This level shows the midbrain region and the tentorium cerebelli. The presence of contrast highlights vascular structures. The petrous part of the temporal bone is visible. 3. Right Image (Non-Contrast Axial CT Scan). This slice provides a view of the lateral ventricles, basal ganglia, and cortical structures. No contrast is administered in this slice, highlighting the normal brain parenchyma.These images together help in the evaluation for various intracranial pathologies like infarctions, hemorrhages, tumors, or other abnormalities.

MLLM-Reformatted ☺

The provided image displays three axial slices of a non contrast-enhanced cranial CT scan. From left to right: The first image shows the cerebellar region, where no signs of cerebellar hemorrhage can be observed, suggesting complete resorption. The tissue density appears normal without any hyperintense areas, which typically signify acute bleeding. The second image focuses closer on the brainstem and cerebellum, also showing no evidence of swelling, and the structures of the posterior fossa appear well-preserved and normal in size. The third image is an axial slice at the level of the lateral ventricles, presenting clear and symmetric ventricles, indicating the absence of hydrocephalus. The fourth ventricle is visible and appears normal, suggesting a regression of previously noted hydrocephalus.

Figure 1: Constructing image captions in various approaches. Detailed explanations of these methods are given in Appendix F. We use *gpt-4* as the LLM and *gpt-4V* as the MLLM. ~~Strikethrough texts~~ indicate erroneous descriptions or descriptions unrelated to the image. This case is sourced from one of the PubMed papers at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2852039/.

Case Analysis Figure 1 presents examples generated by these methods. It shows that Native-Caption is ambiguous and contains content unrelated to the image. LLM-Reformatted misinterprets three sub-images as a CT slide, leading to misleading descriptions, and fails to exclude irrelevant content. GPT4v-Distill generates factually incorrect descriptions due to the lack of contextual text. In contrast, MLLM-Reformatted produces superior descriptions by leveraging both visual information and contextual cues. It accurately and thoroughly describes the key information of the image. The subsequent experiment in Section 4.3 further demonstrates the higher data quality of MLLM-Reformatted.

3 PubMedVision

3.1 Data Collection

To acquire a comprehensive dataset of PubMed medical images, we integrated previously compiled public data of PubMed images, specifically LLaVA-Med PMC (514K) (Li et al., 2023a), PMC-Inline (11M) (Wu et al., 2023), and PMC-OA (1M) (Lin et al., 2023). Although extensive, the majority of this merged data consists of charts and graphs from papers rather than medical images. Therefore, we implemented a rigorous data filtering pipeline: **(1) Text Filtering.** A medical vocabulary was used to filter out data where the contextal text contains a sufficient number of medical terms. **(2) Image Filtering.** We excluded low-resolution images (less

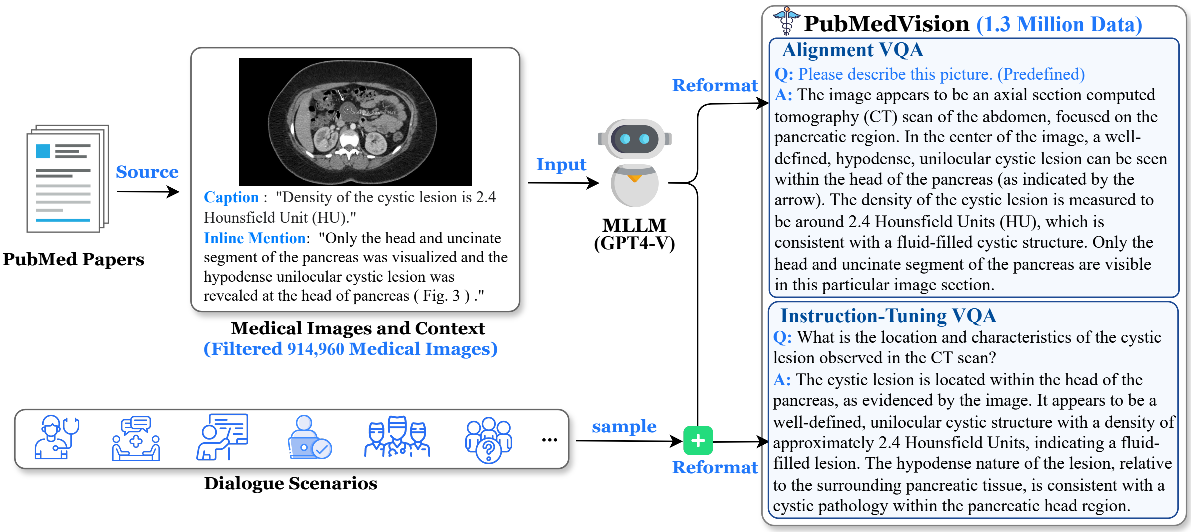


Figure 2: Construction process of the PubMedVision dataset.

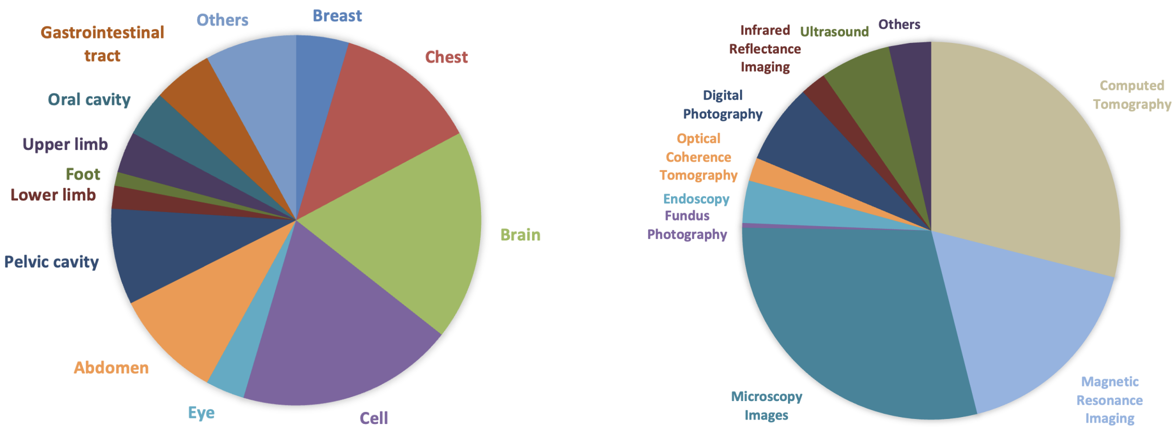


Figure 3: Image Diversity in PubMedVision. A random sample of 500 images from PubMedVision are categorized. **Left:** Distribution of body parts depicted in the images. **Right:** Distribution of imaging modalities.

than 336x336 pixels). A medical image classification model, trained on 1K manually labeled images and 10K MLLM-labeled images, is used to identify medical images. **(3) Deduplication.** Using Sentence-BERT (Reimers and Gurevych, 2019) as the encoder, we obtained semantic embeddings of the image captions and filtered out images with overly similar contexts. For more details, please see Appendix B.

Ultimately, we filtered out 914,960 medical images and their associated contextual text (captions and inline mentions). Figure 3 illustrates the diversity of medical modalities and image regions covered by PubMedVision's images. These medical images are then used to sequentially construct 1.3 million VQA data points for medical alignment.

3.2 MLLMs Reformatting

Each collected data point includes one or more medical images $\mathcal{I}$ and their corresponding contextual image descriptions X. As shown in Figure 2, we provided $\mathcal{I}$ and X to MLLMs to generate medical VQA data. According to ALLaVA (Chen et al., 2024), we generate two types of VQA data to enhance image alignment. Using the prompt shown in Figure 4, the MLLM generates an overall image description d, a specific question q about the image, and the corresponding answer a, as follows:

$$d, q, a = \text{MLLMs}(\mathcal{I}, X)$$

Alignment VQA We predefined a question q' and combined it with the image description d to form Alignment VQA (q', a). The predefined question was sampled from a set of predefined ques-

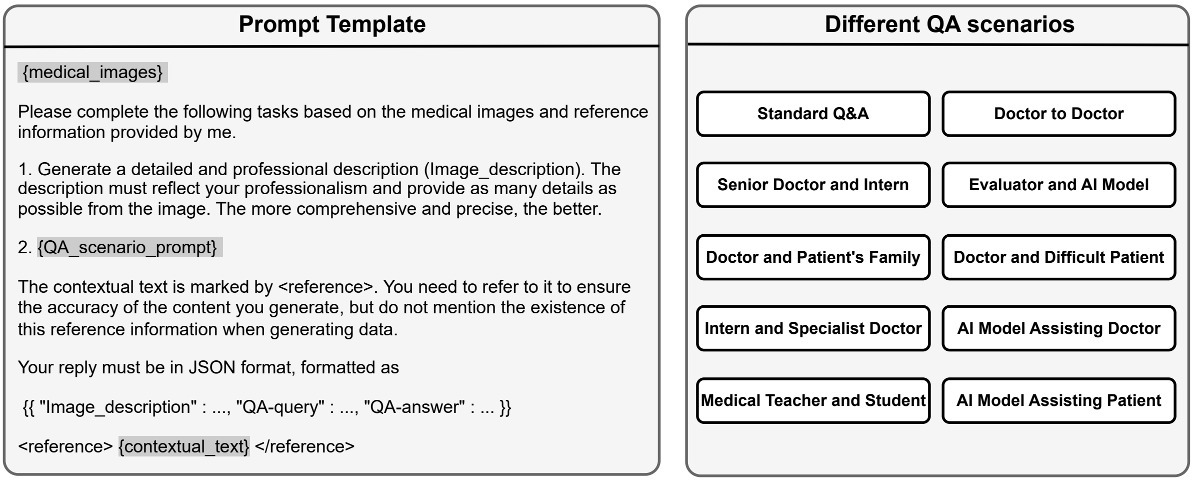


Figure 4: Prompts used for data generation. **{medical_images}** represents medical images. **{QA_scenario_prompt}** denotes scenario prompts, sampled from the scenarios on the right, see Appendix D for details. **{contextual_text}** pertains to image captions and inline mentions.

tions, which can be found in Appendix C. According to ShareGPT-4V (Chen et al., 2023b), such detailed image descriptions help in learning the alignment from image to text.

Instruction-Tuning VQA We used the question q and answer a generated by MLLMs as Instruction-Tuning VQA (q, a) for enhancing instruction-following ability and image comprehension. Unlike Alignment VQA, the questions are generated by MLLMs specifically for the images. To diversify the generated q, we designed eight different scenarios, as detailed in Appendix D. We randomly sample scenario settings into the synthetic prompt to enable MLLMs to generate more varied questions.

Based on this method, we employ GPT4-Vision (*gpt-4-turbo-2024-04-09*) as MLLMs to synthesize 647,031 Alignment VQA and 647,031 Instruction-Tuning VQA. Consequently, PubMedVision contains a total of 1.3 million data points.

4 Experiment

4.1 Experiment Settings

Training and Validation To verify the effectiveness of PubMedVision, we selected the LLaVA-1.5 model architecture combined with LLaMA-3-8B. We use the original settings of LLaVA-1.5, featuring a 336×336 CLIP-Large mode (Radford et al., 2021) and a two-layer MLP Projector. For the base LLM, we utilize LLaMA-3-8B, which is pre-trained on OpenHermes (Teknium, 2023) text instruction data. We followed the same two-stage training method as LLaVA-1.5 (Liu et al., 2024) (Pretraining and Finetuning) and the same hyperparameters (including a learning rate of 2e-5 and one epoch). Based on this setup, we train the following three comparative models:

- **LLaVA-v1.5-LLaMA3-8B** The baseline model that only uses LLaVA-1.5 data. The data distribution is Pretraining: **558K** (LLaVA); Finetuning: **658K** (LLaVA).
- **LLaVA-v1.5-LLaMA3-8B + LLaVA_Med** This model uses both LLaVA-1.5 data and LLaVA_Med's two-stage data. The data distribution is Pretraining: **558K** (LLaVA) + **457K** (LLaVA_Med Alignment); Finetuning: **658K** (LLaVA) + **57K** (LLaVA_Med VQA).
- **LLaVA-v1.5-LLaMA3-8B + PubMedVision** This model uses both LLaVA-1.5 data and PubMedVision data. The data distribution is Pretraining: **558K** (LLaVA) + **647K** (PubMedVision Alignment VQA); Finetuning: **658K** (LLaVA) + **647K** (PubMedVision Instruction-Tuning VQA).

Baselines We compared two types of open-source models: (1) Medical MLLMs. We evaluated three Medical MLLMs, including Med-Flamingo (Moor et al., 2023), RadFM (Wu et al., 2023), and LLaVA-Med-7B (Li et al., 2023a). (2) General MLLMs. We compared the latest models in the LLaVA series, including LLaVA-v1.6-7B, LLaVA-v1.6-13B, and LLaVA-v1.6-34B (Liu et al., 2023). Additionally, we included comparisons with Yi-

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.
Med-Flamingo	45.4	43.5	54.7	23.3	41.7
RadFM	50.6	34.6	38.7	25.9	37.5
LLaVA-Med-7B	51.4	48.6	56.8	24.7	45.4
Qwen-VL-Chat	47.0	56.0	55.1	36.6	48.9
Yi-VL-34B	53.0	58.9	47.3	39.5	49.7
LLaVA-v1.6-7B	52.6	57.9	47.9	35.5	48.5
LLaVA-v1.6-13B	55.8	58.9	51.9	36.6	50.8
LLaVA-v1.6-34B	58.6	67.3	59.1	44.4	57.4
Our Training					
LLaVA-v1.5-LLaMA3-8B	54.2	59.4	54.1	36.4	51.0
+ LLaVA_Med	60.2	61.2	54.5	46.6	55.6
+ PubMedVision	**63.8**	**74.5**	**59.9**	**52.7**	**62.7**
HuatuoGPT-Vision-34B	**68.1**	**76.9**	**63.5**	**58.2**	**66.7**

Table 2: The results of the medical VQA benchmark.

Model	CT	FP	MRI	OCT	Der	Mic	X-Ray	US	Avg.
Med-Flamingo	34.6	33.3	27.5	26.0	28.3	28.1	30.1	33.2	30.2
RadFM	33.3	35.0	22.0	31.3	36.3	28.0	31.5	26.1	30.5
LLaVA-Med-7B	25.3	48.4	35.9	42.1	45.2	44.0	31.7	83.7	44.5
Qwen-VL-Chat	51.5	45.4	43.9	54.0	55.4	49.5	63.1	33.5	49.5
Yi-VL-34B	39.8	57.2	51.4	70.5	54.5	61.4	64.2	40.5	54.9
LLaVA-v1.6-7B	40.1	39.5	54.8	58.4	54.0	48.8	53.3	47.9	49.6
LLaVA-v1.6-13B	40.0	43.6	47.4	63.2	58.0	50.5	59.6	42.6	50.6
LLaVA-v1.6-34B	50.6	63.4	60.9	68.4	65.7	62.8	74.7	44.5	61.4
Our Training									
LLaVA-v1.5-LLaMA3-8B	33.0	49.7	53.8	76.0	63.1	48.4	56.6	31.2	48.8
+ LLaVA_Med	60.8	68.5	**66.3**	79.0	66.6	60.3	73.3	49.3	65.5
+ PubMedVision	**61.6**	**80.2**	65.1	**86.3**	**71.6**	**67.4**	**81.4**	**87.4**	**75.1**
HuatuoGPT-Vision-34B	60.8	**85.5**	**66.5**	**90.0**	**74.0**	**71.3**	**83.8**	81.7	**76.7**

Table 3: The accuracy of OmniMedVQA within different modalities. Specifically, **FP** denotes *Fundus Photography*, **IRI** denotes *Infrared Reflectance Imaging*, **MRI** denotes *Magnetic Resonance Imaging*, **OCT** denotes *Optical Coherence Tomography*, **Der** denotes *Dermoscopy*, End denotes *Endoscopy*, **Mic** denotes *Microscopy Images*, **US** denotes *Ultrasound*.

VL-34B (Young et al., 2024) and Qwen-VL-Chat (Bai et al., 2023).

HuatuoGPT-Vision Building on PubMedVision, we developed our specialized medical MLLM, HuatuoGPT-Vision. It enhances *LLaVA-v1.5-LLaMA3-8B* + *PubMedVision* by featuring: (1) a larger model, utilizing Yi-1.5-34B (Young et al., 2024) as the foundational LLM; (2) bilingual capabilities, supported by an additional 348K *Chinese* medical VQA dataset translated from PubMedVision; and (3) enhanced medical knowledge, with added training from the medical text corpus of HuatuoGPT-II (Chen et al., 2023a).

Benchmarks To verify the medical multimodal capabilities of MLLMs, we employed three types of benchmarks: (1) Medical VQA Benchmark, for which we used the test sets of VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), and PMC-VQA (Zhang et al., 2023c) to assess medical question-answering capabilities. Specifically, for VQA-RAD and SLAKE, we used English CLOSED segment and the accuracy metric. (2) Multimodal Benchmark: MMMU (Yue et al., 2024) is a popular multimodal benchmark, and we utilized the Health & Medicine track of MMMU, which is relevant to medical multimodality. (3) Traditional Medical Imaging Tasks. We used the open access part of the OmniMedVQA dataset (Hu et al., 2024), which includes 42 traditional medical imaging datasets, all formatted as VQA. Note that for all benchmarks, we use the **zero-shot** method and the question template set by LLaVA, as shown in Appendix E.

4.2 Effectiveness of PubMedVision

Medical VQA Benchmarks Table 2 presents the results of the medical VQA benchmarks. General-purpose MLLMs, such as LLaVA-v1.6, demonstrate superior performance compared to medical-

Model	BMS	CM	DLM	P	PH	MMMU Health & Medicine
Med-Flamingo	29.6	28.1	24.8	25.3	31.2	28.3
RadFM	27.5	26.8	25.8	24.7	29.1	27.0
LLaVA-Med-7B	39.9	39.1	34.6	37.4	34.0	36.9
Qwen-VL-Chat	36.5	31.7	32.7	28.4	34.6	32.7
Yi-VL-34B	49.4	48.9	43.2	40.5	32.0	41.5
LLaVA-v1.6-7B	40.5	36.9	32.1	32.3	26.9	33.1
LLaVA-v1.6-13B	53.6	46.7	33.3	22.2	40.0	39.3
LLaVA-v1.6-34B	56.4	56.0	46.9	**46.7**	**41.7**	48.8
Our Training						
LLaVA-v1.5-LLaMA3-8B	42.3	44.0	37.0	34.7	35.2	38.2
+ LLaVA_Med	48.2	43.8	42.0	39.7	35.8	41.1
+ PubMedVision	**61.0**	**58.8**	**50.0**	44.7	38.7	**49.1**
HuatuoGPT-Vision-34B	**64.6**	**62.5**	**50.6**	**54.1**	**44.2**	**54.4**

Table 4: Results on the **test set** for the MMMU Health & Medicine track. The Health & Medicine track is divided into five categories: **BMS** for *Basic Medical Science*, **CM** for *Clinical Medicine*, **DLM** for *Diagnostics and Laboratory Medicine*, **P** for *Pharmacy*, and **PH** for *Public Health*. Results are obtained by submitting to the official website.

specific MLLMs like LLaVA-Med-7B, aligning with the findings of studies (Hu et al., 2024). However, the addition of medical multimodal data to LLaVA-v1.5-LLaMA3-8B significantly enhances performance, revealing substantial potential for improving medical image understanding. Notably, the use of the PubMedVision led to an **11.7%** increase in overall accuracy, significantly outperforming the earlier LLaVA_Med dataset. Additionally, as detailed in Appendix A, fine-tuning on the training sets of these four datasets indicates that PubMedVision can also significantly improves performance in downstream medical multimodal tasks.

OmniMedVQA Evaluation OmniMedVQA integrates 41 traditional medical imaging tasks, all formatted as VQA. Table 3 presents the results of it across 8 different modalities. After incorporating PubMedVision, the performance of LLaVA-v1.5-LLaMA3-8B showed a significant improvement of **26.3%**, which is notably higher than the 16.7% improvement achieved with the LLaVA_Med dataset. With PubMedVision, LLaVA-v1.5-LLaMA3-8B outperforms previous open-source models.

MMMU Health & Medicine Track MMMU is a widely recognized multimodal benchmark, and we utilize its Health & Medicine Track for assessment. Figure Table 4 presents the results of the MMMU test set, showing that LLaVA-v1.5-LLaMA3-8B + PubMedVision surpassed other models in the Health & Medicine Track, with performance comparable to the larger-parameter LLaVA-v1.6-34B. These findings further validate PubMedVision's effectiveness in aligning medical images.

4.3 Data Quality of PubMedVision

Experimental Setup To validate the effect of the MLLM reformatter in PubMedVision, we constructed four datasets based on the four caption construction methods described in Section 2.2. Specifically, we randomly sampled 60,000 image-context pairs from PubMedVision to create these four distinct datasets. For each caption, we pre-set the question: *"Please provide a description of the given medical image"* to form VQA datasets, which we refer to as Native-Captions-60K, LLM-Reformatted-60K, GPT4v-Distill-60K and MLLM-Reformatted-60K. Detailed explanations of these four methods are provided in Appendix F.

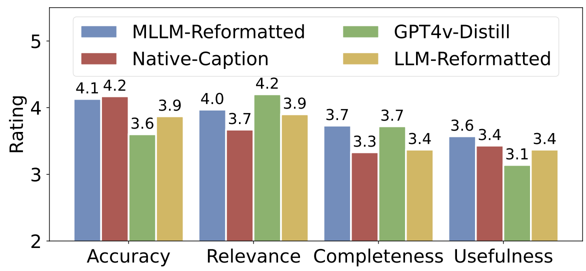


Figure 5: Scoring results from medical experts. Four scoring metrics are detailed in Appendix I.

Expert Evaluation To assess data quality, we randomly sampled 90 images, each contain four descriptions from four data categories, totaling 360

	VQA-RAD	SLAKE	PathVQA	PMC-VQA
LLaVA-v1.5-LLaMA3-8B	54.2	59.4	54.1	36.4
+ Native-Caption-60K	53.5	58.9	52.8	36.9
+ LLM-Rephrase-60K	56.5	63.7	54.0	39.1
+ GPT4v-Distill-60K	55.0	60.6	54.7	35.3
+ PubMedVision-60K	**56.8**	**64.1**	**55.1**	**40.8**
+ Native Caption of PubMedVision	60.8	65.2	56.9	45.6
+ PubMedVision	**63.8**	**74.5**	**59.9**	**52.7**

Table 5: Comparison of different datasets. The **60K** dataset is added only in the second stage of training. **Native Caption of PubMedVision** refers to using the original image captions, incorporated in both phases to match the training of **PubMedVision**.

entries. Three medical experts are invited to evaluate these image descriptions, each reviewing an equal number from each category. The criteria included: 1) Accuracy: correctness of the description, 2) Relevance: relevance to the image and avoidance of irrelevant details, 3) Completeness: inclusion of key medical features, and 4) Usefulness: utility for medical decision-making, diagnosis, and treatment planning. Each item is rated on a scale of 1-5. Detailed scoring criteria are in Appendix I. Table 5 shows the scoring results (average values). Although *Native-Captions* had high accuracy, its relevance and completeness were lacking. *LLM-Reformatted* improved in relevance but still lacked completeness. *GPT4v-Distill* had good relevance and completeness but fell short in accuracy and usefulness. *MLLM-Reformatted* showed the best completeness and usefulness, with high accuracy and relevance, indicating superior quality.

Empirical Evaluation Using LLaVA-v1.5-LLaMA3-8B, we evaluated four datasets to enhance medical multimodal capabilities. As shown in Figure 5, the *MLLM-Reformatted* method outperforms other datasets with the same data volume, demonstrating superior alignment in medical multimodal applications. Additionally, a comparison between the full datasets of PubMedVision and native caption reveals that PubMedVision performs significantly better, supporting the use of MLLMs for data reformatting.

5 Related Works

Multimodal Large Language Models Recent advancements in MLLMs leverage the capabilities of LLMs such as LLaMA to integrate visual features into the textual space. Notably, Flamingo (Alayrac et al., 2022) introduces visual features by incorporating cross-attention layers into LLMs. To align multimodal features effectively, BLIP2 (Li et al., 2023b) integrates a pre-trained visual encoder with LLMs through a novel Q-former. InstructBLIP (Dai et al., 2024) further refines this approach by enhancing performance using instruction-following data. Following this trend, LLaVA (Liu et al., 2024) and subsequent MLLMs (Zhu et al., 2023; Ye et al., 2023) utilize high-quality multimodal data for instruction tuning, demonstrating significant improvements. Additionally, ALLVA (Chen et al., 2024) shows that even a small model (3B) can achieve impressive results with high-quality Visual Question Answering (VQA) data. This underscores the importance of multimodal data.

Medical MLLMs Encouraged by the success of medical LLMs such as ChatDoctor (Yunxiang et al., 2023), MedicalGPT (Xu, 2023), HuatuoGPT (Zhang et al., 2023a; Chen et al., 2023a), and Apollo (Wang et al., 2024), researchers have been focusing on developing a medical Multimodal LLM capable of understanding medical images. Med-Flamingo (Moor et al., 2023) extends Flamingo to the medical domain by utilizing medical multimodal data for pre-training. LLaVA-Med (Li et al., 2023a) refines this approach by filtering image-text pairs from PubMed papers and smaller VQA datasets synthesized by LLMs to train a medical MLLM based on LLaVA's parameters. Additionally, (Zhang et al., 2023c) created the PMC-VQA dataset for medical VQA by self-instruction on PMC-OA (Lin et al., 2023). Using this dataset, they developed MedVInT. RadFM (Wu et al., 2023) integrates a large amount of medical multimodal data, including 2D and 3D radiology images, to construct a radiology MLLM. However, according to recent findings (Hu et al., 2024), current medical models still lag behind general medical models in medical multimodal, indicating that higher quality datasets are needed for medical multimodal applications.

Medical VQA Datasets To enhance image-text alignment and develop medical multimodal chatbots, researchers have focused on constructing medical VQA datasets. VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and Path-VQA (He et al., 2020) are among the earliest medical VQA datasets. However, their sample sizes are small (less than 20K) and their diversity is limited, primarily to radiology modalities. Subsequently, PMC-VQA (Zhang et al., 2023c) expands the dataset scale by using image-text data from PubMed papers and rewriting it into VQA format using LLMs. LLaVA-Med VQA (Li et al., 2023a) data is derived from filtering higher quality data from PMC-15M (Zhang et al., 2023b) and synthesizing VQA using LLMs. PMC-CaseReport (Lau et al., 2018) filters case images from PubMed and generates VQA using LLMs, though it retains only radiology modality images. Currently, there is still a need for more comprehensive and larger-scale medical VQA datasets.

6 Conclusion

In this study, we refined high-quality data from numerous medical image-text pairs on PubMed. We then employ MLLM-powered reformatting method to enhance this data. In this way, we construct PubMedVision, a large-scale, high-quality medical multimodal dataset. Experimental results show that PubMedVision significantly boosts the multimodal capabilities of MLLMs, with marked improvements on benchmarks. This inspires the idea that PubMed holds great potential to advance medical multimodal capabilities, with the key challenge being how to improve data quality, despite the presence of many non-medical images and poor descriptions. We hope that the proposed PubMedVision dataset can aid the development of medical MLLMs in the future.

Limitations

The PubMedVision dataset, while valuable, has several limitations that should be considered:

- **Hallucination of MLLMs**: The construction of the PubMedVision dataset utilizes MLLM models (GPT-4V), which, as generative models, can produce hallucinations or inaccuracies, leading to potential errors in the dataset. Future studies may benefit from enhanced validation processes to mitigate this issue.
- **Limited Scenario Diversity**: The Instruction-Tuning VQA of PubMedVision are generated based on 10 predefined scenarios. This limited scope may have constrained the diversity of the dataset. Expanding the range of scenarios could enhance the dataset's comprehensiveness and applicability to diverse medical situations.
- **Data Selection Bias**: The rigorous image selection strategy during data preparation ensured high-quality data but may have excluded potentially valuable data. Future data collection efforts should consider a more inclusive selection approach to optimize data utility and address potential biases inherent in PubMed data.
- **Limitations of AI-Generated Data**: The reliance on AI-generated content for medical applications, as seen in this dataset, brings inherent limitations and risks, particularly in terms of accuracy and reliability. This should be carefully considered when applying such data in medical contexts.
- **Coverage of Medical Specialties**: There may be gaps in the medical specialties or imaging modalities covered by the dataset, which could limit its effectiveness in addressing certain medical queries or conditions.

Ethical Statement

Our dataset was generated by the GPT4-V model, it may contain hallucinations or inaccuracies. Given this potential limitation, we strictly limit the use of the dataset to research purposes only. It is not to be employed in clinical or other industry applications where its use could lead to unintended consequences due to these possible inaccuracies. We emphasize the ethical responsibility of users to adhere to this restriction to ensure the safety and integrity of their applications.

Acknowledgements

This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing.

References

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

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

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

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

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. 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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. 2024. Reformatted alignment. *arXiv preprint arXiv:2402.12219*.

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

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *arXiv preprint arXiv:2402.09181*.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, et al. 2024. Hidden flaws behind expert-level accuracy of gpt-4 vision in medicine. *arXiv preprint arXiv:2401.08396*.

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

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. 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.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. 2023. Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *arXiv preprint arXiv:2310.19061*.

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

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. 2023b. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A More Experiments

Fine-tuned Results of VQA Benchmarks To verify whether PubMedVision can enhance downstream tasks, we fine-tuned the model using the training set of the Benchmarks. As shown in Table 6, PubMedVision effectively improves downstream medical tasks, significantly benefiting all four VQA downstream tasks. For VQA-RAD and SLAKE, we increased the number of training epochs, and it performed better than traditional fine-tuning methods, such as M3AE (Chen et al., 2022) and RAMM (Yuan et al., 2023). The results are shown in Table 7.

Results on validation set of MMMU Table 8 presents the validation results of MMMU, where LLaVA-v1.6-34B exhibits superior overall performance. However, compared to the test set results of MMMU (official submission) in Table 4, LLaVA-v1.5-LLaMA3-8B combined with PubMedVision demonstrates better performance. Overall, PubMedVision allows the 8B version of LLaVA to achieve effects comparable to the 34B version in medical applications.

Applicability of PubMedVision To verify the applicability of PubMedVision across different MLLM models, we further trained PubMedVision on other MLLM models, specifically LLaVA-v1.5-7B and Qwen-VL-Chat. As demonstrated in Table 9, PubMedVision effectively enhances the medical multimodal capabilities of these diverse MLLM models as well.

B Data Pipline

To acquire a comprehensive dataset of PubMed images, we integrated previously compiled PubMed image and contextual text data, specifically LLaVA-Med PMC data (514K) (Li et al., 2023a), PMC-Inline (11M) (Lau et al., 2018), and PMC-OA (1M) (Lin et al., 2023). Although the dataset is extensive, most of the data consists of charts and graphs from papers rather than medical images. Therefore, we need to filter out higher-quality medical image-text data. We established a pipeline as follows:

1. **Contextual Text Filtering:** Utilizing the SPECIALIST Lexicon [2] from the Unified Medical Language System, we employed GPT-4 to filter out common phrases, creating a refined medical lexicon. Using this lexicon, we assessed the number of medical terms in image captions, filtering out data with fewer than five medical terms. This ensures the captions are sufficiently informative.

2. **Image Filtering:** Initially, we excluded images with a resolution lower than 336x336 pixels to ensure quality. Next, we filtered out chart images to retain only medical images. To accurately identify non-medical images, we manually labeled 1K images and synthesized 10K image labels using MLLMs (GPT4-Vision). We then trained a classifier based on the CLIP image encoder, achieving a 91% accuracy on the validation set. This classifier is used to filter out non-medical images.

3. **Deduplication:** We applied a semantic retriever for deduplication. Using all-mpnet-base-v2 (Reimers and Gurevych, 2019) as the encoder, we generated semantic embeddings of the image captions. We then removed images with an embedding dot product similarity exceeding 480, ensuring a unique and high-quality dataset.

[2] https://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_001.html

C Question Set of Alignment VQA

Alignment VQA is based on the generated image description d and the question q' sampled from a predefined question set. q' is sampled from the multi-image question set (Figure 8) if multiple images are involved, and from the single-image question set (Figure 7) otherwise.

D Prompts for different QA scenarios

In our study, Instruction-Tuning VQA is generated based on ten pre-set different scenarios. This approach covers a broader range of medical topics and scenarios, thereby enhancing the diversity of the QA pairs, and more comprehensively improving the ability to follow instructions. The sampling method also prevents the overconcentration or absence of certain scenarios, contributing to data balance, which in turn improves the performance and stability of the model. Our ten scenario prompts are displayed from Figure 9 to Figure 18.

E Prompts for Evaluation

During the evaluation, we utilized a standardized template, which is illustrated in Figure 6.

Model	VQA-RAD (Finetuned)	SLAKE (Finetuned)	PathVQA (Finetuned)	PMC-VQA (Finetuned)	Avg.
Fine-tuning on the training set.					
LLaVA-v1.5-LLaMA3-8B	63.3	68.9	85.2	50.3	66.9
+ LLaVA_Med	66.3	69.5	90.7	52.7	69.8
+ PubMedVision	**68.9**	**84.1**	**93.0**	**57.3**	**75.8**

Table 6: Results on VQA Benchmarks after fine-tuning on the task training sets. All datasets were trained using their respective in-built training sets, over **2** training epochs.

Model	Fine-tuning Epochs	VQA-RAD	SLAKE
M3AE	50 (VQA-RAD); 15 (SLAKE)	83.5	87.8
RAMM	100	85.3	91.6
LLaVA-8B + PubMedVision (Zero-shot)	0	63.8	74.5
LLaVA-8B + PubMedVision	2	68.9	84.1
LLaVA-8B + PubMedVision	9	86.8	93.5

Table 7: Comparison of model performance fine-tuning on VQA-RAD and SLAKE training sets with different epochs.

Model	BMS	CM	DLM	P	PH	MMMU Health & Medicine
Med-Flamingo	33.6	30.2	23.3	29.3	25.8	28.4
RadFM	31.6	28.2	26.7	26.2	26.8	27.9
LLaVA-Med-7B	50.0	33.3	26.7	40.7	43.3	38.6
Qwen-VL-Chat	39.3	36.7	20.0	29.6	33.3	31.7
Yi-VL-34B	48.1	55.6	36.7	48.1	53.3	48.2
LLaVA-v1.6-7B	46.4	33.3	30.0	29.6	26.7	33.1
LLaVA-v1.6-13B	53.6	46.7	33.3	22.2	40.0	39.3
LLaVA-v1.6-34B	**57.1**	**63.3**	**50.0**	44.4	**63.3**	**55.9**
Our Training						
LLaVA-v1.5-LLaMA3-8B	42.9	43.3	30.0	25.9	50.0	38.6
+ LLaVA_Med	42.9	46.7	36.7	40.7	46.7	42.8
+ PubMedVision	50.0	**63.3**	36.7	**48.1**	53.3	50.3
HuatuoGPT-Vision-34B	**64.3**	60.0	46.7	**66.7**	56.7	**58.6**

Table 8: Results on the **validation set** of MMMU Health & Medicine track. The Health & Medicine track is divided into five categories: **BMS** for *Basic Medical Science*, **CM** for *Clinical Medicine*, **DLM** for *Diagnostics and Laboratory Medicine*, **P** for *Pharmacy*, and **PH** for *Public Health*.

	VQA-RAD	SLAKE	PathVQA	PMC-VQA
LLaVA-v1.5-7B(Liu et al., 2024)	50.6	53.4	52.3	33.1
⊕ PubMedVision	57.5 +6.9	57.6 +4.3	57.6 +4.3	46.3 +13.2
Qwen-VL-Chat(Bai et al., 2023)	47.0	56.0	55.1	36.6
⊕ PubMedVision	54.3 +7.3	66.7 +10.7	57.0 +1.9	48.4 +11.8

Table 9: PubMedVision for other MLLMs, where ⊕ denotes further training with PubMedVision.

F Comparison of Methods for Constructing Multimodal Datasets

Table 12 presents four methods of synthesizing multimodal data. To facilitate a better comparison, we uniformly construct captions using these four methods. These captions are then combined with the query "Please provide a description of the given medical image" to form a VQA dataset for comparing the differences among the various methods.

G Different MLLMs as Reformatters

We use GPT-4V as the data reformatter and compare it with another advanced MLLM, Claude 3.5 Sonnet. The focus is on the quality of generated data, specifically the accuracy and completeness of synthesized captions.

Criterion	Claude-3.5-Sonnet Win	Tie	GPT-4V Win
Accuracy	8	33	11
Completeness	23	11	16

Table 10: Expert evaluation of caption synthesis by Claude 3.5 Sonnet and GPT-4V.

The evaluation was conducted by domain experts who provided a win-tie-loss assessment based on the mentioned criteria. The results are summarized in Table 10. The two models performed similarly: Claude 3.5 Sonnet generated richer descriptions, while GPT-4V was slightly more accurate. Combining their strengths could yield better results, a direction for future research.

H Discrepancies in Evaluation Results with LLaVA-Med

Our evaluation of LLaVA-Med-7B shows some differences from the original paper, as detailed in Table 11.

Using the official inference code, we confirmed no issues with our implementation. The discrepancy likely arises from differences in task prompts. While we used the multiple-choice prompts specified for LLaVA, these may not match those in the LLaVA-Med evaluations, which were not fully described.

Although VQA-RAD and SLAKE scores were lower, PathVQA improved. For fairness, we also compared LLaVA-8B + LLaVA-Med, which used the same two-stage training data. These results were closer to the original, with notable improvement in PubMedVision.

I Scoring Guidelines

In the Expert Evaluation section, the detailed criteria used by experts to score the data are shown in Figure 20.

J Training Resources

We utilized the training setup of LLaVA. For training resources, we used two NVIDIA machines, each equipped with 8 A800 GPUs with 80GB of memory.

K Case Study

We have presented seven samples from the PubMedVision dataset, coming from various scenarios and including both single and multiple figures. These examples are displayed in Table 13 through Table 19.

Model	VQA-RAD	SLAKE	PathVQA
LLaVA-Med (Original)	61.4	52.2	54.1
LLaVA-Med (Our Test)	51.4	48.6	56.8
LLaVA-8B + LLaVA-Med	60.2	61.2	54.5
LLaVA-8B + PubMedVision	63.8	74.5	59.9

Table 11: Comparison of LLaVA-Med Results.

Prompt for Evaluation

```
<question>
A. <option_1>
B. <option_2>
C. <option_3>
D. <option_4>
Answer with the option's letter from the given choices directly.
```

Figure 6: Prompt for Evaluation.

Single-Image Question Set

- `Please describe this picture.`
- `Can you describe the image for me?`
- `What details stand out in this image?`
- `Could you provide a detailed description of what is shown in the picture?`
- `What is the main focus of this photograph?`
- `Describe the composition and the subjects in this picture.`
- `Explain the visual content of the image`
- `Analyze the image in a comprehensive and detailed manner.`
- `Write a detailed description of the given image.`
- `What is this photo about?`
- `What is depicted in the image?`

Figure 7: Single-image question set for alignment VQA. They convey the same meaning using different natural language expressions.

Dataset	Data Synthesis Method
Native-Caption-60K	Uses the native contextual text (Caption and inline Mention) as the image caption.
LLM-Reformatted-60K	Following the synthesis method of LLaVA-Med with LLMs, we provide only the contextual text to LLM (gpt-4-turbo-2024-04-09) to construct answers. For specific prompts, see Figure 19.
GPT4v-Distill-60K	We only provide the image to GPT-4 Vision (gpt-4-turbo-2024-04-09) to generate a description in response to the query "Please provide a description of the given medical image".
MLLM-Reformatted-60K	The method of PubMedVision, where MLLMs construct data based on contextual text and visual information from the image. We use the answers from PubMedVision's Alignment VQA as the constructed caption.

Table 12: Description of four methods for constructing image captions.

Multi-Image Question Set

- Please describe these pictures.
- Can you describe the images for me?
- What details stand out in these images?
- Could you provide a detailed description of what is shown in the pictures?
- What are the main focuses of these photographs?
- Describe the composition and the subjects in these pictures.
- Explain the visual content of the images.
- Analyze the images in a comprehensive and detailed manner.
- Write a detailed description of the given images.
- What are these photos about?
- What is depicted in the images?

Figure 8: Multi-image question set for alignment VQA. They convey the same meaning using different natural language expressions.

Standard Q&A

You need to generate a question-and-answer pair based on this image. The question should be designed to test other models' understanding of this medical image; it should be phrased simply and conversationally. However, your response should be professional, showcasing your understanding of the medical image by providing useful information derived from the image and detailed analysis. The reply should offer detailed and rich useful information.

Figure 9: Prompt for Standard Q&A Scenario: A guide for crafting a standard question-and-answer scenario.

Doctor and Patient's Family

You need to generate a question-and-answer pair based on this image. You need to play the roles of a doctor and a patient's family member, discussing the results shown in the image. The doctor should explain the imaging findings in layman's terms and answer any questions posed by the family member. The family member may inquire about the cause of the disease, severity, treatment options, and related content. The doctor should answer patiently to ensure that the family member fully understands the condition.

Figure 10: Prompt for Doctor and Patient's Family Scenario: A concerned family member inquiring about a patient's condition from the doctor.

Doctor to Doctor

You need to generate a question-and-answer pair based on this image. This pair should be a professional discussion between doctors about the image. You need to mimic a doctor's tone in asking and answering questions. The response should provide detailed and rich useful information derived from the image.

Figure 11: Prompt for Doctor to Doctor Senario: A professional discussion scenario between doctors regarding a medical image.

Intern and Specialist Doctor

```
You need to generate a question-and-answer pair based on this image. You should adopt the tone
of an intern to ask questions and a specialist doctor to answer them. The answers should provide
useful information derived from the image and give a detailed analysis. The response should
provide detailed and rich useful information.
```

Figure 12: Prompt for Intern and Specialist Doctor Scenario: A simulated dialogue where an intern asks questions and a specialist provides detailed, informative answers based on a medical image.

Medical Teacher and Student

```
You need to generate a question-and-answer pair based on this image. You need to act as a medical
teacher and a student, engaging in an educational interaction about the image. The teacher should
pose questions, asking the student to analyze the image and propose possible diagnoses. The
student should answer the questions and explain their observations and reasoning process.
```

Figure 13: Prompt for Medical Teacher and Student Scenario: A simulated educational interaction where the teacher prompts the student to analyze a medical image and propose potential diagnoses.

Senior Doctor and Intern

```
You need to generate a question-and-answer pair based on this image. You should act as a senior
doctor and an intern, discussing the image. The senior doctor should pose relevant questions
to test the intern's observational and analytical skills concerning the image, while the intern
should respond and explain their viewpoint.
```

Figure 14: Prompt for Senior Doctor and Intern Scenario: A simulated dialogue where a senior doctor tests an intern's observational and analytical skills through questions based on a medical image.

Doctor and Difficult Patient

```
You need to generate a question-and-answer pair based on this image. You need to act as a doctor
communicating with a patient who is skeptical about their diagnosis. The patient may pose a series
of tricky questions, questioning the doctor's explanations and treatment suggestions. The doctor
needs to use the imaging data patiently and explain the condition in an easy-to-understand manner,
addressing all the patient's queries to alleviate their concerns and build trust.
```

Figure 15: Prompt for Doctor and Difficult Patient Scenario: A simulated dialogue where a doctor patiently communicates a diagnosis to a skeptical patient, using the image data to explain the condition in a comprehensible way, and address all queries to build trust.

Evaluator and AI Model

```
You need to generate a question-and-answer pair based on this image. You need to act as a member
of a quality control team, focusing on assessing an AI model's visual capabilities in handling
complex medical images. Team members should inquire about subtle details in the image.
```

Figure 16: Prompt for Evaluator and AI Model Scenario: A simulated interaction where a quality control team member assesses an AI model's ability to analyze complex medical images.

AI Model Assisting Doctor

```
You need to generate a question-and-answer pair based on this image. You need to act as a doctor
using an AI model to analyze a medical image to better understand a patient's condition. The
doctor should ask specific questions about structures, abnormalities, and potential clinical
significance visible on the image. The AI model should provide detailed analyses based on its
algorithms but not make final clinical diagnoses. The doctor will use the information provided by
the AI model to aid their diagnostic decision-making process.
```

Figure 17: Prompt for AI Model Assisting Doctor Scenario: A simulated dialogue where a doctor consults an AI model about details in a medical image to improve diagnostic accuracy.

AI Model Assisting Patient

```
You need to generate a question-and-answer pair based on this image. You need to act as an AI model
interacting with a patient who has questions about visible content on their medical image. The
patient may be curious or confused about certain structures or markings on the image and seeks
clear explanations. The AI model should explain specific details such as tissue density, shape,
or any abnormal areas' potential meanings, maintaining simplicity and avoiding excessive medical
jargon. The AI model's response should aim to provide educational information to help the patient
better understand their imaging results, emphasizing that final interpretations and diagnoses
must be done by a professional doctor.
```

Figure 18: Prompt for AI Model Assisting Patient Scenario: A simulated dialogue where an AI model explains details on a patient's medical image, aiming to clarify patient queries, while emphasizing that final interpretations are by professional doctors.

Prompt for LLM-Reformatted

```
You have been provided with textual context information of images from a biomedical research
paper, but you do not have access to the actual image. You need to respond to the following
question based on this image's context information.

In your response, avoid using phrases like 'mentioned', 'caption', or 'context'. Instead,
describe the information as if it were directly observed 'in the image'. Answer responsibly,
avoiding any overconfidence, and refrain from giving medical advice or diagnostic information.
Encourage the user to consult a healthcare professional for further advice.

<Image Context Information>: {image_context_information}

<Question>: Please provide a description of the given medical image.

Please respond to the <Question> as instructed.
```

Figure 19: Prompt for LLM-Reformatted. {image_context_information} pertains to image captions and inline mentions.

Dataset Scoring Guidelines

```
Please rate the image based on the following criteria (1 to 5):

1. Accuracy: The rating should be based on whether the description accurately reflects
the medical features and information in the image. A score of 5 indicates complete accuracy,
while a score of 1 indicates complete inaccuracy.

2. Relevance: The rating should be based on the degree of tight association between
the described content and the key medical features and information in the image, and whether it
successfully avoids the depiction of irrelevant details in the image. A score of 5 indicates
high relevance, while a score of 1 indicates complete irrelevance.

3. Completeness: The rating should be based on whether the description includes all
the key medical features and information in the image. A score of 5 indicates complete
completeness, while a score of 1 indicates significant omissions.

4. Practicality: The rating should be based on the extent to which the description
helps with medical decision-making, diagnosis, and treatment planning. A score of 5 indicates
high practicality with in-depth descriptions, while a score of 1 indicates that the description
is superficial and not practical.
```

Figure 20: Dataset Scoring Guidelines.

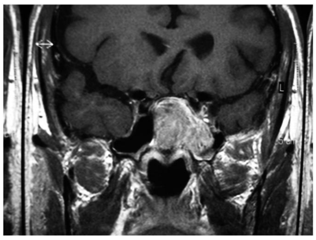
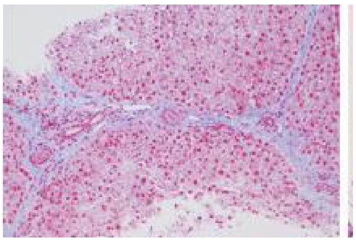
Sample 1: Standard Q&A	
	
Description	The image shows a T1-weighted MRI scan with intravenous gadolinium enhancement. The scan reveals a well-defined, homogeneously enhancing mass located within the left sphenoid sinus. The mass appears to be occupying a significant portion of the sphenoid sinus, with no obvious signs of infiltration into adjacent structures. The surrounding soft tissues appear normal, and there is no evidence of bony erosion or destruction. Overall, the imaging findings are consistent with a sphenoid sinus mass, likely a benign tumor or other pathological process confined within the sphenoid sinus.
QA-query	What is the location of the mass observed in the MRI image?
QA-answer	The mass observed in the MRI image is located within the left sphenoid sinus. The sphenoid sinus is one of the paranasal sinuses, located deep within the skull base, behind the eyes. The MRI scan clearly shows the mass occupying a significant portion of the left sphenoid sinus, with no obvious extension into surrounding structures.

Table 13: Sample 1 for Standard Q&A Senorio.

Sample 2: Evaluator and AI Model



Description	This image depicts a microscopic view of a liver tissue section stained with trichrome, highlighting various histological features. The predominant colors are varying shades of pink and red, representing hepatocytes, and blue areas indicating connective tissue and collagen fibers. The structure of the tissue is organized with distinct lobular architecture. The hepatocytes are mostly uniform in size and appearance, with some regions showing slight signs of ballooning, indicative of early hepatocyte degeneration which is not overtly prominent. Sinusoids are visible between hepatocyte plates, and there are no clear signs of inflammation or fibrosis within the tissue section viewed.
QA-query	Given the subtle indications of hepatocyte changes in the image, can you identify any specific patterns or regions that may suggest early stages of ballooning degeneration?
QA-answer	Yes, some hepatocytes in the image show slight enlargement and pale cytoplasm which could be indicative of early stages of ballooning degeneration. These changes are subtle and not uniformly spread across the entire section, appearing more as isolated phenomena rather than widespread pathological changes.

Table 14: Sample 2 for Evaluator and AI Model Senorio.

Sample 3: Intern and Specialist Doctor	
	
Description	The image presents a cross-sectional view from a chest computed tomography (CT) scan, showcasing the right lung of the patient. Prominently visible is a solid mass with spiculated edges located in Segment 9b of the right lung. This mass exhibits a defined shape with a maximum diameter of approximately 30mm, indicating its noticeable size relative to surrounding lung structures. Additionally, the lung parenchyma displays scattered areas of decreased attenuation consistent with emphysematous changes, which are indicative of air trapping and lung tissue degradation.
QA-query	As an intern, I noticed a solid mass with spiculated edges on this CT scan. Could you explain the potential implications of these features and the significance of its location in the right lung's Segment 9b?
QA-answer	This CT scan reveals a solid mass with spiculated margins located in Segment 9b of the right lung, which significantly raises concerns for a malignant tumor, possibly a type of non-small cell lung cancer. Spiculated edges are typically suggestive of an aggressive tumor that may invade adjacent tissues, a finding that warrants further diagnostic evaluation and possibly biopsy to confirm malignancy. The location in Segment 9b, which is a part of the lung often less affected by other common conditions, emphasizes the need for targeted investigation. Managing such a case would likely involve a multidisciplinary approach including oncology, thoracic surgery, and radiology, to determine the best therapeutic strategies and follow-up care.

Table 15: Sample 3 for Intern and Specialist Doctor Senorio.

Sample 4: Doctor and Difficult Patient	
	
Description	The image is a coronal reformat of an abdominal CT scan. It distinctly displays two major findings. On the left side of the image (patient's right side), a mass is visible in the head of the pancreas, indicated by a thin arrow. This mass appears as a localized area of denser tissue compared to its surroundings. On the right side of the image (patient's left side), there is evidence of massive gastric dilatation, marked by a thick arrow. This dilatation shows the stomach expanded significantly, occupying a large portion of the abdominal cavity, which is abnormally large compared to the usual size of the stomach.
QA-query	Doctor, how serious is the mass in my pancreas and what are the potential risks with this gastric dilatation?
QA-answer	The mass in the head of your pancreas could represent several things ranging from benign conditions to more serious ones like pancreatic cancer. We will need further tests, possibly including a biopsy, to determine the exact nature of the mass. As for the gastric dilatation, it's quite significant and can lead to discomfort, pain, and other problems like nausea or vomiting if not managed timely. It may also indicate an obstruction or a severe motility disorder. Both conditions are serious and require careful evaluation and management. We'll plan some follow-up assessments to understand better and address both the mass and the dilatation appropriately, ensuring the best possible treatment plan is set for you.

Table 16: Sample 4 for Doctor and Difficult Patient Senorio.

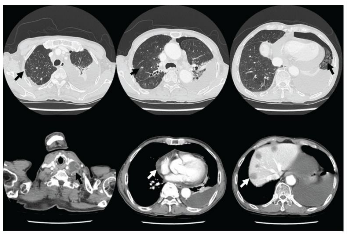
Sample 5: Doctor and Patient's Family	
	
Description	The image displays a series of axial cuts of computed tomography (CT) scans from the thorax and abdomen. In the thoracic region, there are multiple tiny nodules and cystic lesions visible in both lungs. A significant left pleural effusion is evident with signs of pneumothorax, characterized by the presence of air in the pleural space which can be identified by the visceral pleural line and a clear space devoid of lung markings. Additionally, there is evidence of pericardial effusion as seen by the abnormal accumulation of fluid around the heart. The CT scans of the abdomen show enlargement of multiple lymph nodes and multiple hepatic masses, which appear as irregular areas within the liver.
QA-query	Doctor, can you explain what all these findings mean? How serious are these conditions and what are the possible treatments?
QA-answer	Certainly, I understand your concerns. The findings indicate a few serious conditions. Multiple tiny nodules and cystic lesions in the lungs could be indicative of an infectious or inflammatory process, or less likely, metastatic disease. The large amount of left pleural effusion and the presence of pneumothorax are concerning as they can compromise lung function. Pleural effusion means there is fluid accumulation that can restrict lung expansion, and pneumothorax involves air leaking into the space between the lung and chest wall, which can cause the lung to collapse. Pericardial effusion, the fluid around the heart, can affect heart function if severe. The enlarged lymph nodes and hepatic masses in the abdomen might suggest a possible malignancy or metastatic spread. Treatment options would vary based on the underlying cause but could include drainage of fluid from the pleural and pericardial spaces, antibiotics or chemotherapy if infection or malignancy are confirmed, and potentially surgery to address the hepatic masses. A multidisciplinary team approach will be crucial in managing these conditions effectively.

Table 17: Sample 5 for Doctor and Patient's Family Senorio.

Sample 6: Medical Teacher and Student (Multiple Images)

Description	The provided images depict two AP (anteroposterior) pelvis radiographs. The first radiograph shows both hips with symmetric placement of screws or pins in what appears to be a bilateral surgical intervention, possibly for hip stabilization. There is visible deformation of the pelvic bones and the acetabulae (hip sockets) show signs of protrusion into the pelvic cavity. The second image features more pronounced skeletal changes including a notably misaligned right femur with surgical hardware adjustments, possibly indicating a post-injury or post-surgical state with a complex fixation device attached to the right side extending from the pelvis to the femur. These images suggest significant progression in the patient's condition, hinting at ongoing issues with hip stability and pelvic deformation.
QA-query	Given these pelvis radiographs, can you describe any observable changes or indicators that suggest a progression in the patient's pelvic condition?
QA-answer	Certainly, examining the radiographs, one can note an increased deformation in the pelvic region on the second image compared to the first. This includes greater protrusion of acetabular regions and a more complex assembly of surgical hardware on the right side, suggesting increased intervention. Such findings likely indicate a significant progression of underlying conditions affecting the pelvic structure which necessitated additional surgical measures over time.

Table 18: Sample 6 for Medical Teacher and Student Senorio. (Multiple Images)

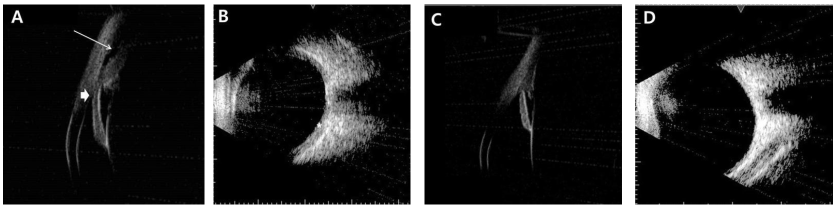
Sample 7: Evaluator and AI Model (Multiple Images)	
	
Description	These images are medical ultrasound scans of the eye. Image A shows an ultrasound biomicroscopy (UBM) scan depicting angle closure with a shallow anterior chamber (indicated by the thick arrow) and anterior rotation of the ciliary process with ciliochoroidal effusion (thin arrow). Image B is a B-scan ultrasonography showing diffuse thickening of the choroid (asterisk). Image C, taken one month after treatment, shows a UBM scan of the left eye with a normal anterior chamber angle and the disappearance of ciliochoroidal effusion. Image D is another B-scan ultrasonography, which shows no choroidal thickening.
QA-query	In Image B, can you identify the specific location and extent of the choroidal thickening observed in the B-scan ultrasonography?
QA-answer	In Image B, the B-scan ultrasonography shows diffuse thickening of the choroid, as indicated by the asterisk. The thickening appears to be widespread throughout the choroidal region, extending across a significant portion of the eye's posterior segment.

Table 19: Sample 7 for Evaluator and AI Model Senorio. (Multiple Images)