SemiHVision: Enhancing Medical Multimodal Models with a Semi-Human Annotated Dataset and Fine-Tuned Instruction Generation

Anonymous ACL submission

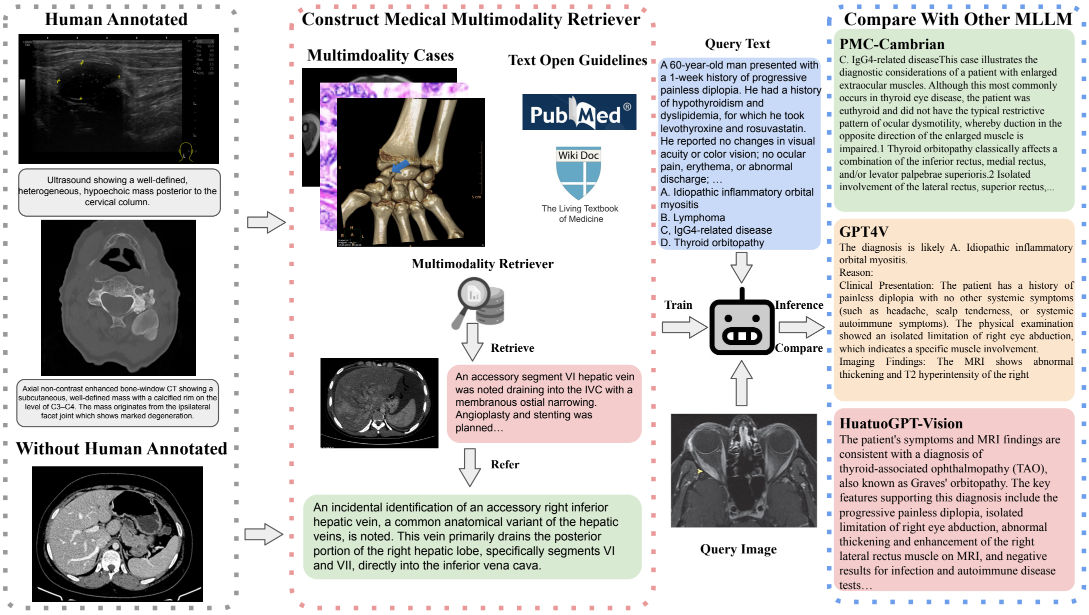
Abstract

Multimodal large language models (MLLMs) have made significant strides, yet they face challenges in the medical domain due to limited specialized knowledge. While recent medical MLLMs demonstrate strong performance in lab settings, they often struggle in real-world applications, highlighting a substantial gap between research and practice. In this paper, we seek to address this gap at various stages of the end-to-end learning pipeline, including data collection, model fine-tuning, and evaluation. At the data collection stage, we introduce SemiHVision, a dataset that combines human annotations with automated augmentation techniques to improve both medical knowledge representation and diagnostic reasoning. For model fine-tuning, we trained PMC-Cambrian-8B-AN over 2400 H100 GPU hours, resulting in performance that surpasses public medical models like HuatuoGPT-Vision-34B (79.0% vs. 66.7%) and private general models like Claude3-Opus (55.7%) on traditional benchmarks such as SLAKE and VQA-RAD. In the evaluation phase, we observed that traditional benchmarks cannot accurately reflect realistic clinical task capabilities. To overcome this limitation and provide more targeted guidance for model evaluation, we introduce the JAMA Clinical Challenge, a novel benchmark specifically designed to evaluate diagnostic reasoning. On this benchmark, PMC-Cambrian-AN achieves state-of-the-art performance with a GPT-4 score of 1.29, significantly outperforming HuatuoGPT-Vision-34B (1.13) and Claude3-Opus (1.17), demonstrating its superior diagnostic reasoning abilities.

1 Introduction

Multimodal foundation models have demonstrated remarkable success across a wide range of applications by integrating visual and textual information, showcasing their ability to process complex visual patterns alongside natural language (Yan et al., 2023; Liu et al., 2024b; Jin et al., 2024). This success has led to increasing interest in applying these models to medical tasks that involve both medical images and text-based descriptions. Recent advances have focused on fine-tuning general multimodal models on medical datasets composed of image-text pairs, yielding promising results (Li et al., 2024; Chen et al., 2024b).

However, despite these advancements, Multimodal Large Language Models (MLLMs), such as Claude3 and GPT-4V, face significant challenges in the medical domain due to their limited ability to understand domain-specific visual features. Unlike general tasks, medical image interpretation requires both the identification and understanding of the semantics of an image, including anatomical landmarks, and expert medical knowledge, which are crucial for accurate diagnosis—a level of complexity not typically required in general vision-language tasks. Furthermore, medical imaging spans multiple modalities—such as X-ray, CT, MRI, and DSA—each requiring specialized knowledge for proper interpretation. For instance, white regions in CT scans and MRI images convey entirely different meanings, underscoring the need for modality-specific experts. General models like GPT-4V lack comprehensive medical knowledge, further limiting their effectiveness in such specialized applications. Compounding these challenges, obtaining high-quality annotated medical data is particularly difficult due to privacy concerns and the significant costs of expert annotation (Xie et al., 2024; Bustos et al., 2020; Lau et al., 2018; Irvin et al., 2019; Johnson et al., 2019; Ikezogwo et al., 2024). These factors limit the scalability and performance of MLLMs in medical applications, highlighting the urgent need for more robust, domain-adapted models capable of handling the unique complexities of medical multimodal tasks. Therefore, some researchers have explored the use of synthetic medical data to fine-



Ultrasound showing a well-defined, heterogeneous, hypoechoic mass posterior to the cervical column.

Axial non-contrast enhanced bone-window CT showing a subcutaneous, well-defined mass with a calcified rim on the level of C3–C4. The mass originates from the ipsilateral facet joint which shows marked degeneration.

An accessory segment VI hepatic vein was noted draining into the IVC with a membranous ostial narrowing. Angioplasty and stenting was planned…

An incidental identification of an accessory right inferior hepatic vein, a common anatomical variant of the hepatic veins, is noted. This vein primarily drains the posterior portion of the right hepatic lobe, specifically segments VI and VII, directly into the inferior vena cava.

A 60-year-old man presented with a 1-week history of progressive painless diplopia. He had a history of hypothyroidism and dyslipidemia, for which he took levothyroxine and rosuvastatin. He reported no changes in visual acuity or color vision; no ocular pain, erythema, or abnormal discharge; …
A. Idiopathic inflammatory orbital myositis
B. Lymphoma
C, IgG4-related disease
D. Thyroid orbitopathy

PMC-Cambrian
C. IgG4-related diseaseThis case illustrates the diagnostic considerations of a patient with enlarged extraocular muscles. Although this most commonly occurs in thyroid eye disease, the patient was euthyroid and did not have the typical restrictive pattern of ocular dysmotility, whereby duction in the opposite direction of the enlarged muscle is impaired.1 Thyroid orbitopathy classically affects a combination of the inferior rectus, medial rectus, and/or levator palpebrae superioris.2 Isolated involvement of the lateral rectus, superior rectus,...

GPT4V
The diagnosis is likely A. Idiopathic inflammatory orbital myositis.
Reason:
Clinical Presentation: The patient has a history of painless diplopia with no other systemic symptoms (such as headache, scalp tenderness, or systemic autoimmune symptoms). The physical examination showed an isolated limitation of right eye abduction, which indicates a specific muscle involvement.
Imaging Findings: The MRI shows abnormal thickening and T2 hyperintensity of the right

HuatuoGPT-Vision
The patient's symptoms and MRI findings are consistent with a diagnosis of thyroid-associated ophthalmopathy (TAO), also known as Graves' orbitopathy. The key features supporting this diagnosis include the progressive painless diplopia, isolated limitation of right eye abduction, abnormal thickening and enhancement of the right lateral rectus muscle on MRI, and negative results for infection and autoimmune disease tests…

Figure 1: Our pipeline starts with two types of data: human-annotated and unannotated medical images. For the human-annotated dataset, we employ GPT-4o to generate instruction-based QA pairs and reformat the existing captions. In parallel, a multimodal retriever constructs a corpus by indexing data from OpenGuidelines (Chen et al., 2023) and the augmented dataset. For the unannotated dataset, the system retrieves relevant guidelines or similar cases, providing them as context to GPT-4o for generating instructions and augmented captions. Finally, we benchmark our model's performance against HuatuoGPT-Vision and GPT-4V, demonstrating its enhanced reasoning and captioning capabilities.

tune general MLLMs into medical MLLMs, such as LLaVA-Med (Li et al., 2024) and HuatuoGPT Vision (Chen et al., 2024b), which have surpassed general-domain MLLMs on traditional benchmark datasets.

Current medical datasets, particularly synthesis datasets, are often limited by the absence of detailed manual annotations, relying heavily on model-generated descriptions that fail to adequately integrate local and global image information. We found that the quality of synthesis data is concerning, as it lacks manual annotations and is dependent on general MLLMs that do not possess sufficient medical knowledge. This shortcoming leads to increased hallucinations and diminished performance (Pal et al., 2023). Additionally, we found that most existing medical benchmark datasets, such as LLaVA-Med-VQA (Li et al., 2024) and PubmedVision (Chen et al., 2024b), exhibit poor quality and lack inference capacities, particularly in tasks that require specialized expertise. These datasets, which are primarily focused on PubMed image-caption pairs, also lack reasoning and diagnostic reasoning datasets, resulting in the trained models lacking reasoning capabilities (Dorfner et al., 2024). In contrast, general domain models tend to perform well in reasoning but lack the medical domain knowledge necessary for producing accurate and clinically relevant outcomes (Yuksekgonul et al., 2023; Geirhos et al., 2020). In fact, there has been ongoing debate regarding whether medical LLMs or general LLMs perform better in medical tasks (Lehman et al., 2023; Dorfner et al., 2024). Therefore, this paper will also explore this discussion.

Furthermore, the quality of existing benchmarks raises concerns. For instance, datasets like SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), and Path-VQA (He et al., 2020) focus heavily on knowledge recall rather than assessing how to use medical knowledge for inference. Our experiments reveal that these benchmarks employ a limited evaluation methodology that fails to adequately assess the reasoning capabilities of the models. Consequently, researchers tend to develop medical LLMs in a direction that prioritizes these benchmarks, leading to a paradox where medical MLLMs perform exceptionally well on these benchmarks yet may not exhibit strong performance in real-world medical tasks. This situation calls into question the effectiveness of current evaluation methods in guiding medical MLLMs toward improved real-world clinical performance. Here we propose three key concerns: 1) Do medical MLLMs actually outperform general MLLMs in clinical tasks, as indicated by traditional benchmarks? 2) Do current evaluation methods effectively guide medical MLLMs toward enhancing real-world clinical performance? 3) How can we train a medical MLLM with robust diagnostic capabilities?

Here, we hypothesize that improved benchmark datasets help enhance medical MLLMs, and we have built a comprehensive evaluation pipeline. To this end, we deploy an expert-in-the-loop pipeline to construct high-quality benchmark datasets to train medical MLLMs for knowledge inference. This involves conducting comprehensive evaluations using both traditional benchmarks and the JAMA Clinical Challenges benchmark, a real-world clinical challenge dataset focusing specifically on fine-grained reasoning and diagnostic tasks. We then design novel evaluation metrics to assess medical MLLMs and compare them with general MLLMs. Recognizing the limitations of current multimodal models, we develop **SemiHVision**, a dataset that combines human annotations with automated augmentation techniques. This dataset is constructed using a multimodal retriever, UniIR, which retrieves relevant medical guidelines based on image content and integrates human-labeled regions of interest (ROIs) to guide the model in understanding critical image areas. Lastly, to address the need for medical MLLMs with strong diagnostic abilities, we train **PMC-Cambrian-AN**, first pretraining it on 14 million image-text pairs from the PubMed dataset and then fine-tuning it on the SemiHVision dataset. This results in a model that excels in both knowledge retention and diagnostic reasoning, as demonstrated by its superior performance on the JAMA Clinical Challenges benchmark. Overall, this comprehensive approach—through new datasets, benchmarks, and evaluation pipelines—ensures that medical MLLMs are not only effective in traditional tasks but also better equipped for real-world clinical applications, outperforming both public medical and general MLLMs.

- We design an expert-annotation-in-the-loop pipeline to generate SemiHVision as shown in Figure 1[1].
- We train PMC-Cambrian-AN on PubMed and SemiHVision dataset to get a robust medical inference capabilities MLLM.
- *It is commonly believed that models excelling on traditional benchmarks will also perform well in clinical tasks, with medical MLLMs expected to outperform general ones.* We propose a new evaluation pipeline to compare PMC-Cambrian-AN with other medical and general MLLMs on both traditional and new benchmarks questions this assumption and provides a more accurate evaluation of clinical relevance.

2 Related Work

2.1 Existing Multimodal Medical Datasets

The construction of comprehensive medical multimodal datasets has garnered significant attention. Previous efforts have primarily focused on collecting images paired with clinical reports from specialists, which provide detailed descriptions, including disease types and reasoning. However, many of these datasets have significant limitations. For instance, MIMIC-CXR-JPG (Johnson et al., 2019) consists of 227,835 lung CT images, offering valuable insights but limited in broader medical applications. PMC-OA (Lin et al., 2023) attempts to address scalability with 1.65 million image-caption pairs from the PMC dataset; however, the lack of detailed human-annotated captions for subfigures results in lower quality of information. Datasets such as PMC-CaseReport (Wu et al., 2023), PMC-VQA (Zhang et al., 2023), and LLaVA-Med VQA (Li et al., 2024) and PubMedVision (Chen et al., 2024b) focus on unbalanced modalities and body parts, further restricting applicability. RadGenome-Chest CT (Zhang et al., 2024) includes comprehensive annotations but still relies heavily on paired image-text data, limiting scalability. Early datasets like VQA-RAD, SLAKE, and Path-VQA are constrained by small size and narrow focus. MedTrinity (Xie et al., 2024), although featuring multiple modalities and detailed annotations, relies on Med-LLaVA (Touvron et al., 2023) for text generation, increasing the risk of dataset hallucinations. Additionally, the dataset's questions are narrow in scope, leading to a lack of diversity in QA pairs for instruction tuning. In contrast, our work addresses these challenges by constructing a large-scale medical dataset that includes diverse modalities such as X-ray, CT, and MRI, integrating human annotations, medical guidelines retrieved by a multimodal retriever, and GPT-4o to ensure balanced and comprehensive training across various

---

[1] Case comes from https://www.eurorad.org/case/18708, https://www.eurorad.org/case/17297

medical tasks.

Dataset	Image Retriever	ROI	Human
PMC-CaseReport	×	×	×
PMC-OA	×	×	×
LLaVA-Med VQA	×	×	×
PubMedVision	×	×	×
Medtrinity	×	✓	×
SemiHVision	✓	✓	✓

Table 1: Comparison of Medical Instruction Dataset

2.2 Medical Multimodal Model

In recent years, adapting multimodal foundation models to medical vision-language tasks has gained prominence due to their success in capturing complex visual features (Moor et al., 2023; Li et al., 2024). Current Medical Multimodal Large Language Models (MLLMs) typically pair a visual encoder with a text-only LLM, aligning image data with language understanding. Previous efforts, such as Med-Flamingo (Moor et al., 2023) and Med-PaLM (Tu et al., 2024), fine-tuned general multimodal models on smaller medical datasets, achieving notable results. Med-Flamingo enhanced OpenFlamingo-9B (Chen et al., 2024a) with medical data, while Med-PaLM adapted PaLM-E (Driess et al., 2023) using 1 million data points. Similarly, LLaVA-Med, Med-Gemini (Saab et al., 2024), and HuatuoGPT Vision utilized specialized datasets and instruction tuning to refine medical question-answering tasks. Our work employs Cambrian (Tong et al., 2024), a multimodal large language model with a vision-centric approach, to bridge the gap between visual representation and language understanding. Experimental results show that our model outperforms other state-of-the-art models across multiple tasks.

3 SemiHVision

3.1 Data Collection

Data Source and Image Selection Strategy For the pretraining phase, we utilized the PubMed dataset, initially containing about 25 million samples. After filtering out corrupted or too brief entries (fewer than 20 words), we reduced it to 14 million samples suitable for effective training. In the fine-tuning phase, we incorporated the PMC dataset but faced significant imbalance due to many non-medical images; thus, we employed GPT-4o mini for classification to retain only medical content. Despite these efforts, modalities like MRI and X-ray remained underrepresented. Consequently, we focused on datasets emphasizing CT, X-ray, MRI, histopathology, and pathology to cover a wide range of anatomical regions, prioritizing those with human-annotated information(The details are shown in Appendix A.7 and Table 9). For 3D images, we used annotation information, such as slice IDs, to select images and evenly sampled additional slices to ensure each 3D image corresponded to no more than 20 2D slices. Our experiments indicate that **utilizing all available slices from 3D images results in diminished model performance, further validating our selective approach in the fine-tuning process.**

Human Annotation Preprocessing While many datasets include human annotations like brief reports, the variability in these annotations poses challenges for model training. To address this, we leveraged GPT-4o to regenerate text based on images and annotations, standardizing content into a consistent representation for more effective learning. Additionally, some annotations, such as those in the Eurorad dataset, are notably lengthy, encompassing individual image descriptions, comprehensive findings, and discussions. We segmented the task into three components: generating findings for individual images, consolidating these into overall image findings, and extending them to generate the discussion section(the details are shown in Appendix A.8). Experimental results show that **incorporating these human-annotated datasets enhances the model's fine-grained reasoning capabilities**, as human annotations highlight important details, and GPT-4o's augmentation generates numerous fine-grained reasoning tasks.

3.2 Data Distribution Analysis

Unlike traditional methods for generating instruction datasets, we collected a broader range of human-annotated data across multiple modalities. We conducted a distribution analysis on randomly sampled 200k entries from both the original PMC and SemiHVision datasets. Expert annotators classified the images into categories such as X-ray, DSA, CT, MR, PET/SPECT, Ultrasound, Histopathology, and others. Additionally, we employed GPT-4o for image classification, and to ensure accuracy, a random sample of 100 images was reviewed by human experts, yielding a classification accuracy of 73%. We focused on analyzing higher-frequency modalities, as depicted in Figure 2. The analysis revealed that non-medical images constitute a significant portion of the original PMC dataset, with simulated illustrations like sta-

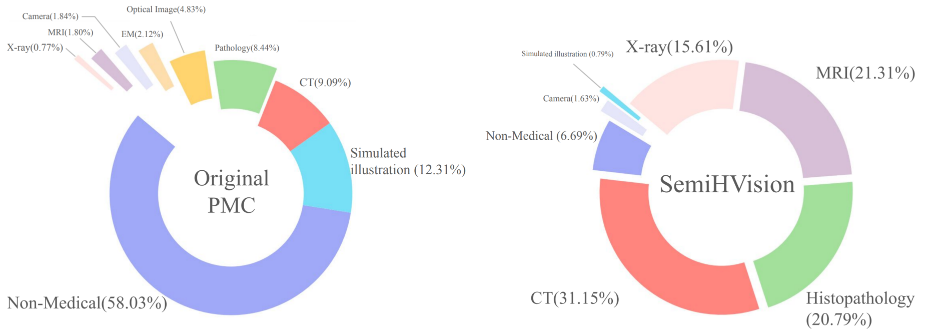


Figure 2: A comparative distribution of image modalities between the original PMC dataset and the SemiHVision dataset. The original PMC dataset contains a significant portion of non-medical content (58.03%), with a relatively lower representation of key medical imaging modalities like MRI (1.80%) and X-ray (0.77%). In contrast, the SemiHVision dataset demonstrates a more balanced distribution, with a substantial increase in clinically relevant modalities such as CT (31.15%), MRI (21.31%), and X-ray (15.61%), while minimizing the presence of non-medical images (6.69%).

tistical charts being the second largest category. In contrast, clinically critical modalities like CT, MRI, and X-ray were significantly underrepresented, highlighting the scarcity of these essential medical images in the PMC dataset. Despite prior filtering efforts, the low representation of modalities like MRI and X-ray means the final dataset still lacks sufficient numbers of these images. For the SemiHVision dataset, we performed a similar sampling and distribution analysis. Unlike the PMC dataset, not all entries were classified using GPT-4o, as some, such as those from Quilt-1M, were already pre-labeled. The resulting distribution demonstrates that SemiHVision contains a more balanced representation of clinically relevant modalities. Notably, modalities underrepresented in the PMC dataset, such as MRI and X-ray, have a much higher proportion in SemiHVision, ensuring more comprehensive coverage of medical knowledge essential for model training and expanding the scope of medical expertise.

3.3 Data Construction Pipeline

Our pipeline comprises three key stages for constructing robust multimodal datasets for medical applications. In the first stage, we develop a multimodal retrieval system by establishing two retrieval databases: a text-based OpenGuidelines repository and a collection of image-text pairs with human annotations from Eurorad and Radiopaedia. Both repositories are indexed for efficient retrieval. For images without human annotations, the retriever takes the image as input to fetch relevant guidelines and analogous cases. In the second stage, we leverage GPT-4o to generate comprehensive clinical reports. For images with existing annotations, we augment these reports with additional context and medical insights; for unannotated images, the retrieved guidelines are used to automatically generate detailed, domain-specific descriptions, ensuring alignment with the medical context and reducing the risk of hallucination. In the final stage, we construct specialized medical question-answer pairs based on both the image-caption pairs and the generated clinical reports. These QA pairs focus on critical diagnostic reasoning and medical decision-making, serving as high-quality instruction tuning data to optimize the model's ability to handle complex medical queries. The entire process is illustrated in Figure 1.

To address the limitations of the general-purpose MLLMs in generating precise medical content, we developed a specialized medical retrieval system. This system utilizes two datasets: the text-based OpenGuidelines and a multimodal clinical case collection from Eurorad and Radiopaedia, covering a wide range of medical fields such as Abdominal Imaging, Uroradiology & Genital Male Imaging, Paediatric Radiology, Neuroradiology, Musculoskeletal System, Interventional Radiology, Head & Neck Imaging, Genital (Female) Imaging, Chest Imaging, Cardiovascular, Breast Imaging, and Hybrid Imaging. For images lacking individual captions but accompanied by an overall "image findings" section, we employed GPT-4o to generate detailed captions for each sub-image based on the case-level descriptions, ensuring every image had a detailed caption without hallucinations. Our retrieval system leverages the UniIR framework and fusion scoring function. In datasets without human annotations, our multimodal retriever fetches four relevant guidelines or cases, including at least one pure text-based guideline for knowledge enrichment. We also incorporated human-annotated regions of interest (ROI) when available to guide GPT-4o in generating precise image captions based on both the ROI and the retrieved corpus. This methodology ensures a comprehensive understanding of each image and its medical context, enhancing both the corpus and retrieval system with domain-specific expertise. We use this multimodal medical knowledge retrieval to generate our instruction dataset (details are provided in Appendix A.1).

4 PMC-Cambrian: Experimental on SemiHVision

4.1 Training Setting

During the training of PMC Cambrian, we employed a two-stage process. First, we filtered the original PMC dataset by removing captions with fewer than 20 words, yielding a final dataset of 14 million samples. We then pre-trained the model on this refined dataset using a learning rate of 1e-4 and an image token length of 512. DeepSpeed Stage 2 was utilized, with a batch size of 8 and a gradient accumulation step of 6. During this stage, we focused solely on training the adapter while freezing the other model components. The pre-training phase ran on four H100 GPUs for 420 hours.

In the fine-tuning phase, we used the SemiHVision dataset with a learning rate of 2e-5, while keeping the DeepSpeed Stage 2 configuration, with a batch size of 6 and a gradient accumulation step of 6. Unlike the pre-training phase, the full model parameters were trained. This fine-tuning process was conducted on 8 H100 GPUs for 90 hours. For instruction tuning, we divided the process into two phases: standard instruction tuning and the Annealing phase which is the same as Llama3 (Dubey et al., 2024). The learning rate in Annealing phase is 1e-5. During the instruction tuning phase, we used non-human-annotated data, primarily GPT-4o-generated synthetic data. In the Annealing phase, we focused on human-annotated data, where GPT-4o applied further augmentation to enhance the dataset(The details are shown in Appendix A.2).

4.2 Automatic Evaluation Pipeline

We evaluate on traditional benchmark and our new benchmark data sets(The details are shown in Appendix A.3). Although several methods exist for measuring textual similarity, such as F1 or ROUGE, both metrics have significant limitations in the medical domain. Therefore, we propose a very strict evaluation pipeine by using two evaluation metrics: the USMLE-Factuality score and the GPT-4o score. For the GPT-4o score, directly allowing GPT-4o to grade the answers is often ineffective, as GPT-4o tends to favor answers that align with its preferred linguistic style, which may not match our intended criteria. Thus, we introduce a scoring framework to evaluate model's fine grained diagnostic ability based on three aspects: **Key Points**, **Inference**, and **Evidence** which is designed by doctors(The details are shown in Appendix A.1).

5 Results

5.1 Traditional Benchmark Result

Table 2 shows that PMC-Cambrian models fine-tuned on GPT-4o synthetic data significantly outperform both general-purpose and medical-specific models across various medical VQA benchmarks. Specifically, PMC-Cambrian-8B achieves an impressive 67.1% average accuracy, surpassing all other tested models. Despite being larger, HuatuoGPT-Vision-34B attains a slightly lower average accuracy of 66.7%, indicating that PMC-Cambrian-8B performs better even with fewer parameters. Compared to similar-sized models, PMC-Cambrian-8B outperforms LLaVA-8B with PubMedVision by 4.4%, highlighting the effectiveness of our high-quality data.

We also tested a variant, PMC-Cambrian-20M, which does not adopt the SemiHVision method. Instead of selecting slices from 3D medical images, all slices were extracted and directly fed into the model for training. Surprisingly, the performance decreased, as reflected in its average accuracy of 63.8%. This decline is attributed to the fact that many of the extracted slices were highly similar to each other. Additionally, several slices did not contain abnormal areas such as tumors, leading to the generation of a large volume of "healthy" data for the model. This oversampling of normal data negatively impacted the model's overall performance.

To demonstrate the importance of annealing, we trained two models: PMC-Cambrian-8B-Mix, which mixes GPT-4o synthetic data and human-annotated data, and PMC-Cambrian-8B-AN, which is first trained on GPT-4o synthetic data and then annealed on human-annotated data. PMC-Cambrian-8B-AN achieves an outstanding 79.0% average accuracy, surpassing PMC-Cambrian-8B-Mix (72.2%) and outperforming HuatuoGPT-

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.
GPT-4o-mini	45.9	59.0	37.9	33.3	44.0
Claude3-Opus	52.5	55.2	54.3	60.7	55.7
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-7B	52.6	57.9	47.9	35.5	48.5
LLaVA-13B	55.8	58.9	51.9	36.6	50.8
LLaVA-34B	58.6	67.3	59.1	44.4	57.4
LLaVA-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
Our Model					
PMC-Cambrian-8B-20M	67.8	76.1	57.8	53.6	63.8
PMC-Cambrian-8B	69.2	77.2	63.6	58.4	67.1
PMC-Cambrian-8B-Mix	**74.2**	**81.3**	**76.3**	**59.1**	**72.2**
PMC-Cambrian-8B-AN	**86.1**	**87.7**	**80.4**	**61.9**	**79.0**

Table 2: Performance comparison of various models on medical VQA benchmarks (VQA-RAD, SLAKE, PathVQA, PMC-VQA) with average scores is presented. PMC-Cambrian-8B-20M refers to the model trained using all slices from the 3D dataset. PMC-Cambrian-8B prioritizes human-annotated slices and selectively sampled portions for training, using GPT-4o-generated synthetic data. PMC-Cambrian-8B-Mix is trained by combining both the human-annotated datasets and the GPT-4o-generated synthetic datasets. PMC-Cambrian-8B-AN is the result after annealing on human-annotated datasets based on PMC-Cambrian-8B.

	Claude3-Opus	GPT-4o-mini	Huatuo-7B	Huatuo-34B	PMC-Cambrian	PMC-Cambrian-AN
Accuracy	58.4	46.2	34.5	44.7	41.2	**58.5**
UMLS Factuality	0.18	0.16	0.13	0.16	0.11	**0.23**
GPT-4 Overall	1.17 ± 0.04	0.91 ± 0.06	1.08 ± 0.03	1.13 ± 0.05	0.78 ± 0.04	**1.29**±0.02
GPT-4 Key-Points	1.27	0.99	1.11	1.01	0.82	**1.28**
GPT-4 Inference	**1.56**	1.13	1.06	1.06	0.63	1.32
GPT-4 Evidence	0.67	0.60	1.08	**1.31**	0.89	1.27

Table 3: UMLS-F and GPT-4 score on JAMA Clinical Challenge across 6 different models :Claude3-Opus, GPT-4o-mini, Huatuo-GPT-Vision 7B, Huatuo-GPT-Vision 34B, PMC-Cambrian, PMC-Cambrian-AN. We also change Deepseek model to evaluate them to eliminate the bias as shown in Table 8

Vision-34B by 18.4%. This performance gap emphasizes the superiority of PMC-Cambrian-8B-AN models, which integrate advanced medical data augmentation and optimization techniques. Compared to private models like Claude3-Opus (55.7%) and GPT-4o-mini (44.0%), PMC-Cambrian-8B-AN models consistently excel across all benchmarks, underscoring the importance of well-curated medical-specific datasets in enhancing multimodal medical understanding.

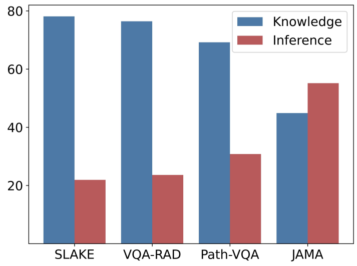


Figure 3: This figure illustrates the proportion of questions assessing knowledge and inference in the Slake, VQA-RAD, Path-VQA, and JAMA Clincial Challenge datasets.

5.2 Diagnostic Benchmark Result

While current public medical multimodal large language models (MLLMs) have demonstrated superior performance over general-domain models in traditional benchmarks—occasionally even surpassing advanced models like Claude3-Opus—a critical question arises: **Do medical MLLMs actually outperform general MLLMs in clinical tasks, as suggested by traditional benchmarks?** To investigate this, we evaluated six models Claude3-Opus, GPT-4o-mini, Huatuo-7B, Huatuo-34B, PMC-Cambrian, and PMC-Cambrian-AN using the JAMA Clinical Challenge dataset and our new evaluation pipeline (see Table 3). We assessed both the accuracy and diagnostic reasoning capabilities of these models. Accuracy was measured using standard methodologies for close-ended QA tasks, while diagnostic reasoning was evaluated through the automatic scoring pipeline described in Section 4.4, which measures performance across three key dimensions: *Key Points*, *Inference*, and *Evidence*. Our results reveal that, although public medical MLLMs perform well on traditional benchmarks (e.g., PMC-Cambrian-AN achieving an accuracy score of 58.5%), they struggle with the JAMA dataset. For instance, Huatuo-34B excels in the Evidence dimension with the highest score of 1.31, surpassing Claude3-Opus's score of 0.67, but exhibits weaker inference capabilities, scoring only 1.06. This suggests that while Huatuo-34B's larger size and medical-specific training enable it to memorize medical knowledge effectively, this does not translate into superior diagnostic reasoning. In contrast, Claude3-Opus, lacking domain-specific medical knowledge, achieves a strong inference score of 1.56, outperforming all other models in this dimension. Additionally, GPT-4o-mini, a general-purpose model, attains a higher inference score (1.13) than Huatuo-34B (1.06), indicating that inference capabilities may not be solely dependent on medical knowledge. Therefore, we conclude that medical MLLMs do not necessarily outperform general MLLMs in clinical tasks requiring diagnostic reasoning.

Furthermore, we also did human evaluation: we engaged three medical professionals to review and assess a small sample of 20 questions drawn from the test sets. These doctors were asked to provide their preference for the rationales generated by two models (Claude3-Opus and SemiHVision) given the gold rationale. The result aligns with our automatic evaluations on the small samples, where PMC-Cambrian-AN achieved a win rate of 0.57 compared to Claude3-Opus which also proved that our automatic evaluation method is reliable.

To address the concern **How can we train a medical MLLM with robust diagnostic capabilities?** We trained PMC-Cambrian using instruction tuning. Initially, PMC-Cambrian was capable of answering medical QA tasks but scored lower across all metrics particularly in inference (0.63) due to its lack of training on human-annotated diagnostic datasets. After applying the annealing process, the enhanced model PMC-Cambrian-AN achieved the highest overall performance, with a top GPT-4 overall score of 1.29. This significant improvement underscores the importance of incorporating human-annotated diagnostic datasets during training, which substantially enhances diagnostic reasoning capabilities. Our findings demonstrate that models like PMC-Cambrian-AN, which integrate high-quality, human-annotated diagnostic data, can outperform models trained solely on synthetic or unannotated data, such as PubMedVision.

Another critical concern is: **Do current evaluation methods effectively guide medical MLLMs toward improving real-world clinical performance?** To investigate this, we classified questions from various datasets into two categories using GPT-4o: **Knowledge-Based Questions** (requiring minimal inference) and **Inference-Based Questions** (requiring reasoning to reach a diagnosis). For PathVQA, GPT-4o's domain-specific knowledge was insufficient, leading to lower classification accuracy. Domain experts classified 100 samples, achieving agreement scores above 0.7 for SLAKE, VQA-RAD, and JAMA, but below 0.6 for PathVQA. To address this, three experts classified the PathVQA dataset, and their annotations were averaged with GPT-4o's output to derive the final distributions in Figure 3. The results show that knowledge-based questions dominate SLAKE (78.1%), VQA-RAD (76.4%), PathVQA (69.2%), and JAMA (44.9%). These findings indicate that traditional benchmarks focus heavily on knowledge recall, while real-world diagnostic tasks, such as those in the JAMA dataset, emphasize inference and reasoning. We conclude that current evaluation methods may not effectively guide medical MLLMs toward improving real-world clinical performance, as they overlook the critical reasoning skills required in clinical settings.

6 Conclusion

In conclusion, this paper highlights the diagnostic and infer ability shortcomings of current medical MLLMs, substantiating these issues through comprehensive experiments. We identify the lack of human-annotated diagnostic datasets as a key reason behind the poor diagnostic performance of medical MLLMs, as many existing datasets rely on GPT-4-generated synthetic data not human annotated diagnosis datasets. To address this, we propose a new instruction-tuning dataset, SemiHVision, and train PMC-Cambrian-AN, which achieves state-of-the-art performance on traditional benchmarks. Furthermore, we introduce the JAMA Clinical Challenge benchmark and a new evaluation pipeline to assess diagnostic reasoning, demonstrating that PMC-Cambrian, trained with SemiHVision, outperforms both public medical MLLMs and private general-domain models like Claude3-Opus in diagnostic tasks.

7 Limitations and Ethical Considerations

Despite the promising results demonstrated by PMC-Cambrian-AN, several limitations warrant consideration. Firstly, the coverage of anatomical regions in our dataset is limited due to the scarcity of high-quality, human-annotated medical data. While we have incorporated multiple imaging modalities such as X-ray, CT, and MRI, the representation across different body parts remains uneven. This imbalance may affect the generalizability of our model in diverse clinical scenarios, potentially limiting its performance on underrepresented regions. Additionally, the model size is constrained to 8 billion parameters, which, while efficient for training and deployment, may restrict the ability to handle more complex reasoning tasks that require deeper understanding and broader context. Exploring larger model architectures could enhance diagnostic performance in future work.

Moreover, the broader societal impacts of deploying PMC-Cambrian-AN necessitate careful consideration. Automated medical systems hold significant potential for improving healthcare efficiency and accuracy but could also influence the roles of medical professionals and patient care practices. It is crucial to approach the implementation of such technological solutions with caution, ensuring they serve as a complement rather than a replacement to the expertise of healthcare professionals. Balancing technological advancement with ethical considerations is essential to maximize benefits while mitigating potential risks in clinical practice.

References

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*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024a. Visual instruction tuning with polite flamingo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17745–17753.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, et al. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

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*.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2024. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

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*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

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.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

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. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. 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.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

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

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. 2024. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

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*.

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*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*.

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

A Appendix

A.1 Template Prompt

Generate Instruction Data In constructing our instruction dataset, we utilize both closed-ended and open-ended question formats. For closed-ended data, such as PMC-VQA, Amboss VQA, JAMA train VQA, Slake train VQA, VQA-RAD train, and Path VQA, we generate answer options only. For open-ended tasks, particularly from JAMA datasets, we also require the model to provide reasoning along with the answers. Additionally, GPT-4o is employed to generate question-answer pairs (QAPs) based on the images and their corresponding augmented captions, with each caption paired with 3 to 10 QAPs depending on its length and complexity. The questions generated are carefully designed to be directly related to the images, ensuring that answers can either be explicitly found or inferred from the caption content. The template prompt deatils are shown in Table 5. This approach minimizes dataset's hallucinations by grounding GPT-4o's output in the information provided in the captions and image data. Furthermore, we utilize a multigranular informtaion, such as specific ROI, and the broader medical context that connects local and global abnormalities to improve model's fine grained ability. By following this structured methodology, we ensure the generation of high-quality, clinically relevant instruction data that improves the accuracy and interpretability of the models.

Evaluation Pipeline Prompt: When evaluating close QA, we only need to calculate accuracy. However, many open QA tasks, such as diagnostic reasoning questions in the JAMA Clinical Challenge, present additional challenges. Although several methods exist for measuring textual similarity, such as F1 or ROUGE, both approaches have significant limitations in the medical domain. Therefore, we propose a very strict evaluation pipeine by using two evaluation metrics: the USMLE-Factuality score and the GPT-4o score. For the GPT-4o score, directly allowing GPT-4o to grade the answers is often ineffective, as GPT-4o tends to favor answers that align with its preferred linguistic style, which may not match our intended criteria. Thus, we introduce a scoring framework to evaluate model's fine grained diagnostic ability based on three aspects: **Key Points**, **Inference**, and **Evidence** which is designed by doctors(The details are shown in Appendix A.1):

- **Key Points** assess whether the model's answer includes the critical elements present in the ground truth.
- **Inference** evaluates whether the diagnostic reasoning in the model's answer is correct, follows the same steps as the ground truth, and whether any key steps are omitted.
- **Evidence** examines whether the model's answer provides the crucial evidence to support its conclusions or diagnostic reasoning.

Finally, an average score will be calculated to represent the overall quality of the answer. To further reduce the influence of linguistic style on GPT-4's scoring, we propose revising all model-generated answers through GPT-4, ensuring that all outputs align with GPT-4's own style distribution. During this revision, GPT-4 will only see the model's answer, without access to any other information.

When scoring, GPT-4 will generate its own summaries of **Key Points**, **Inference**, and **Evidence** based on the ground truth. When assigning scores to these aspects, GPT-4 will no longer see the original answer but will only reference its summarized **Key Points**, **Inference**, and **Evidence**. For further details, please refer to Table 6, 7.

A.2 Instruction Tunning

We employed an annealing strategy in training PMC-Cambrian-AN to enhance its diagnostic capabilities. Empirically, annealing on small amounts of high-quality, human-annotated data significantly boosts performance on key benchmarks. Similar to Llama3, we performed annealing with a data mix that prioritizes high-quality data in select domains, excluding any training sets from commonly used benchmarks. This approach allowed us to assess the true few-shot learning capabilities and out-of-domain generalization of PMC-Cambrian-AN.

We evaluated the efficacy of annealing on the JAMA Clinical Challenge and other diagnostic reasoning benchmarks. The annealing process substantially improved the performance of the pre-trained PMC-Cambrian-8B model, demonstrating enhanced reasoning abilities and clinical applicability. These improvements suggest that, even with a model size constrained to 8 billion parameters, strategic annealing with high-quality data can compensate for limitations in model scale, enabling the model to handle complex reasoning tasks requiring deeper understanding. The whole training phase is shown in figure 4.

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-8B	66.3	69.5	90.7	52.7	69.8
HuatuoGPTVision-8B	68.9	84.1	93.0	57.3	75.8
PMC-Cambrian	88.3	91.1	92.7	88.6	90.2

Table 4: Finetuning results on VQA-RAD, SLAKE, PathVQA, and PMC-VQA datasets.

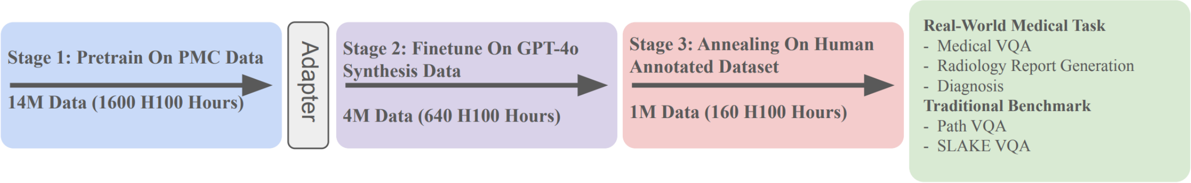

Stage 1: Pretrain On PMC Data
14M Data (1600 H100 Hours)
Adapter
Stage 2: Finetune On GPT-4o Synthesis Data
4M Data (640 H100 Hours)
Stage 3: Annealing On Human Annotated Dataset
1M Data (160 H100 Hours)
Real-World Medical Task
- Medical VQA
- Radiology Report Generation
- Diagnosis
Traditional Benchmark
- Path VQA
- SLAKE VQA

Figure 4: We apply three stages to train PMC-Cambrian.

A.3 Baseline & Benchmark

Medical MLLMs: We evaluated three medical multimodal large language models (MLLMs): Med-Flamingo (Moor et al., 2023), RadFM (Wu et al., 2023), LLaVA-Med-7B (Li et al., 2024) and HuatuoGPTVision-34B (Chen et al., 2024b).

General MLLMs: We assessed the latest models from the LLaVA series, including LLaVA-v1.6-7B, LLaVA-v1.6-13B, and LLaVA-v1.6-34B (Liu et al., 2024a). Additionally, we compared these models with Yi-VL-34B (Young et al., 2024) and Qwen-VL-Chat (Bai et al., 2023). Additionally, we also evaluated several closed-source models: GPT-4-O-Mini and Claude3-Opus.

To evaluate the medical multimodal capabilities of the MLLMs, we employed two types of benchmarks:

Medical VQA Benchmark: We used the test sets from VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), and PMC-VQA (Zhang et al., 2023) to assess the models' medical question-answering abilities. The experiment settings are the same as HuatuoGPT Vision.

New Diagnosis Reason Benchmark Task: To test the model's inference and medical knowledge capabilities, we will evaluate several medical multimodal models on the JAMA Clinical Challenge datasets. The JAMA Clinical Challenge dataset presents complex real-world cases from the Journal of the American Medical Association, challenging models with diagnostic and management tasks based on clinical data and imaging. Together, these datasets provide rigorous benchmarks for assessing the diagnostic and decision-making performance of MLLMs in real-world clinical settings.

A.4 Fine-tuned Results

To assess the impact of SemiHVision on downstream tasks, we applied fine-tuning using the benchmark training sets. As illustrated in Table 4, SemiHVision substantially enhances performance in downstream medical tasks, providing notable improvements across all four VQA tasks.

A.5 Language Style Influence

While our method still utilizes GPT-4o, it effectively eliminates the influence of language style. This is because our scoring is based primarily on whether key points are covered and whether there are any hallucinated key points. Each key point corresponds to a separate score, so variations in language style do not affect the outcome—language style won't cause the model to include more or fewer key points. It's true that switching to a different evaluation model may lead to slight differences in the extracted key points, which could influence the absolute score. However, keep in mind that these key points are derived from the ground-truth answer, and LLMs generally perform very well in summarization tasks. So while there may be changes(for example some model will summarize the most five key points but GPT4o will summarize 10 points), they do not affect the relative ranking of the scores. For fairness, we also evaluated the subset of data using DeepSeek as the scoring model. As shown in the Table 8, although the absolute values differ slightly, the relative scores remain consistent.

A.6 Factuality metrics: UMLS-F1

To evaluate the factual accuracy of LLM outputs, we leverage the UMLS concept overlap metric. The

Unified Medical Language System (UMLS) (Bodenreider, 2004) enhances biomedical interoperability by unifying a comprehensive collection of biomedical terminologies, classification systems, and coding standards. By reconciling semantic variances and representational disparities across different biomedical repositories, UMLS facilitates standardized understanding.

We employ the Scispacy library[2] to identify and align medical named entities in texts with their corresponding UMLS concepts. Scispacy excels in entity recognition, enabling accurate association of named entities in LLM outputs with relevant UMLS concepts, a critical capability for assessing factual accuracy.

Our analytical process utilizes precision and recall metrics. Precision measures the proportion of shared concepts between the LLM output and the ground truth, indicating factual correctness. Recall assesses how well the LLM output covers the concepts present in the ground truth, reflecting the relevance of the information. Formally, given the concept sets from the ground truth (C_{ref}) and the LLM output (C_{gen}), precision and recall are calculated as:

$$\text{Precision} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{gen}}|}, \tag{1}$$

$$\text{Recall} = \frac{|C_{\text{ref}} \cap C_{\text{gen}}|}{|C_{\text{ref}}|}. \tag{2}$$

The F1 score, derived from these precision and recall values, provides a balanced measure of the LLM output's accuracy and relevance.

A.7 Data Source

The fine-tuning datasets include DeepLesion, MIMIC-CXR-JPG, PadChest, Quilt, LLD-MMRI, and MAMA-MIA, along with benchmark training QA datasets such as VQA-RAD, Path VQA, PMC VQA, and Slake VQA, covering multiple modalities like CT, MRI, X-ray and so on. Additionally, we expanded the dataset with data from Eurorad and Radiopaedia to include more diverse modalities as shown in table 9. Additionally, to enable the model to support multiple languages, such as Chinese, we randomly selected 300k datasets and translated them into Chinese for training.

A.8 Human Evaluation and Case Study

Case Study for Evaluation We selected a case from the JAMA Clinical Challenge to evaluate the diagnostic reasoning capabilities of different models, as shown in Table 12[3]. In the case we apply three different colors: red, blue, brown to ask GPT-4O to annotated key points, inference points and evidence points. Our analysis revealed that Claude3-Opus performed accurate inference but lacked detailed evidential support. PMC-Cambrian was able to generate diagnostic reasoning with comprehensive evidence, incorporating most of the important key points. In contrast, HuatuoGPTVision-34B and HuatuoGPTVision-7B failed to capture the essential key points and were unable to effectively utilize medical knowledge for detailed inference, despite having access to extensive medical information that could provide evidence.

Human Annotated Sample Training Data We sampled a case from EURORAD[4]. For EURORAD Dataset, there are serveral sections: Image Caption, Clinical History, Image Findings and Discussion as shown in Table 13. The Image Caption provides a concise description of each image presented. The Clinical History records the patient's medical background and presenting symptoms. In the Imaging Findings section, experts analyze the images to arrive at a diagnostic conclusion, combining observations from all available imaging modalities. The Discussion elaborates on the inference steps and presents the evidence supporting the diagnosis, along with relevant background information to aid in understanding how the conclusion was reached. We also present one sample for our SemiHVision dataset.

Case Study for Multimodality Retriever We did a case study to prove the important of multimodality retriever in our pipeline as shown in Table 14. The inclusion of a retriever in the image description task introduces a marked improvement in the specificity and accuracy of the generated descriptions. Without the retriever, the model (GPT-4o) provides a generalized description of the image, identifying broad anatomical landmarks (heart, aorta, and vertebral column) and speculating on potential abnormalities, such as a mass or vascular anomaly. While the description is coherent, it lacks precision,

---

[2] Using the Scispacy *en_core_sci_lg* model

[3] The case is sourced from https://jamanetwork.com/journals/ jamaophthalmology/fullarticle/2681464.

[4] The case is sourced from https://www.eurorad.org/case/16705.

as the model does not have access to clinical guidelines or related cases, resulting in a speculative rather than a diagnostic interpretation.

In contrast, when the retriever is introduced, the model is supplemented with relevant clinical guidelines and case data, significantly enhancing its diagnostic accuracy. For example, in the case with the retriever, GPT-4o correctly identifies the subaortic ventricular septal defect (VSD) and provides a detailed explanation of its location, dimensions (2.7 cm), and potential clinical implications, such as abnormal blood flow and symptoms like shortness of breath. The addition of retriever-assisted information allows the model to go beyond general observations and offer more specific, clinically relevant insights, directly aligning the image interpretation with known medical cases.

Human Annotator Information We worked with six annotators, all of whom are experts. By experts, we mean either individuals with an MD degree or radiologists with over 10 years of clinical experience. For the image classification task, the three annotators hold MD degrees or work on radiology more than 10 years. For the subsequent human evaluation tasks, such as the one conducted on the JMLR dataset, we engaged three senior radiologists who assessed the model outputs with reference to the ground truth. Each of these doctors has more than ten years of professional experience.

Table 5: Generate Instruction Data Prompt Example Template.

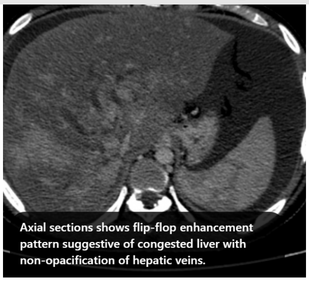
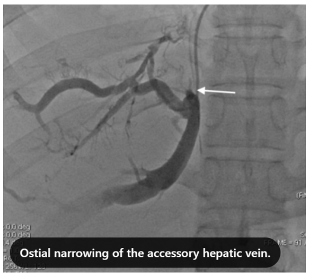
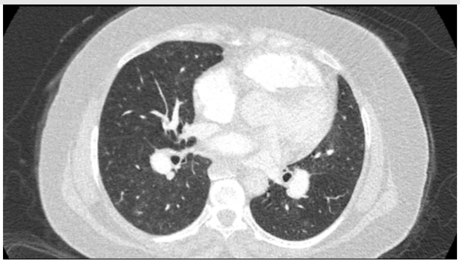
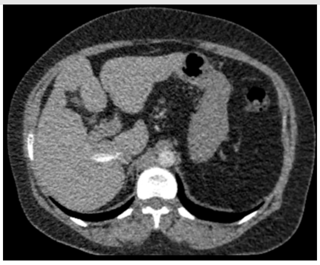
System Prompt	Analyze the provided MRI image and generate a detailed and professional medical report that describes only the abnormalities, significant features, or relevant observations directly seen in the image. Use precise medical terminology and maintain a formal tone. Do not include any introductory phrases, such as "The provided image reveals," or any concluding remarks. Here are some relevant medical guidelines and Clinical cases for you to generate.
Medical Guideline	Angioplasty (PTBA) of the hepatic vein is a safe and effective treatment for Budd-Chiari syndrome (BCS) caused by hepatic venous outflow obstruction. This study, conducted between September 1996 and October 2008, included 101 patients (52 males, 49 females) with a mean age of 31.3 years, all presenting with symptomatic portal hypertension. Of these, 92 patients underwent successful PTBA, targeting the right, left, or accessory hepatic veins, with a technical success rate of 91%. PTBA significantly reduced hepatic venous pressure…
Instruction Prompt	Your second task is to generate 1-2 valuable questions and their corresponding answers that are relevant to the image's content and it would be better that the answers could be explicitly found within the discussion.
Clinical Case	 Axial sections shows flip-flop enhancement pattern suggestive of congested liver with non-opacification of hepatic veins.  Ostial narrowing of the accessory hepatic vein. **Image Findings:** The patient underwent contrast-enhanced computed tomography which showed features of a congested liver with flip-flop pattern of enhancement. Hepatic veins show hypoattenuation on delayed phase. An accessory hepatic vein is also noted in segment VI. A diagnosis of Budd Chiari syndrome (BCS) was made on the basis of the clinical and imaging features. The patient was referred to the interventional radiology team for an endovascular rescue. On conventional venogram, the diagnosis of BCS was confirmed as the hepatic veins were thrombosed. An accessory segment VI hepatic vein was noted draining into the IVC…
Format Prompt	Return the results in the following format: **Report:** report content **Question:**Question content **Answer:**Answer content. Don't generate any other information Here is the image and discussion:
Title: Accessory right inferior hepatic vein	  **Discussion:**Marked dilatation of the pulmonary trunk (6.7 cm) with the right (5.4 cm) and left (4 cm) main branches. Lung window shows mild bilateral diffuse faint groundglass centrilobular lung nodules that may reflect an underlying infection. Scans through the upper abdomen revealed average size cirrhotic liver and reflux of contrast into the IVC and hepatic veins with Incidental opacification of accessory right inferior hepatic vein...

Table 6: Evaluation Pipeline Prompt Example Template.

Extract Key Points	Based on the question and answer, summarize ten key points that you consider to be the most crucial from the standard answer. Return the response in the following format: {1.2.3....} Here is the question:{question} Here is the answer:{answer} Please do not provide any additional information.
Key Points	1. Multifocal electroretinogram (ERG) showed reduced signal in the right eye throughout the macula, confirming the diagnosis of AZOOR.2. Acute zonal occult outer retinopathy (AZOOR) was first described by Gass in 1993...
Extract Diagnostic Reasoning	Based on the question and answer, please provide a detailed summary of the diagnostic reasoning from the standard answer. Return the response in the following format: {1.2.3....} Here is the question:{question} Here is the answer:{answer} Please do not provide any additional information.
Diagnostic Reasoning	1. The patient is a 7-year-old boy with a slowly growing, asymptomatic lump on the left lower neck since birth.2. Physical examination showed a yellowish, hump-like mass with a hairy surface and cartilage-like consistency near the left sternocleidomastoid muscle...
Extract Evidence	Based on the question and answer, please provide a detailed evidence list which is proposed by correct answer. Return the response in the following format: {1.2.3....} Here is the question:{question} Here is the answer:{answer} Please do not provide any additional information.
Evidence	1. Slowly growing, asymptomatic lump on left lower neck since birth.2. Physical examination revealed a yellowish, hump-like mass with hairy surface and cartilage-like consistency.3. Ultrasonography indicated a hypoechoic, avascular, bulging nodule with an anechoic tubular structure.4. MRI demonstrated a protuberant nodule with diffuse...
Key Points Score	Act as a USMLE evaluator, your role involves assessing and comparing a medical student's explanation to the provided target answer. Begin the assessment by carefully reviewing the provided target answer. Then, based on following specific criteria, determine the score for the student's answer. Please judge whether medical student's answer include these key points(or some other relevant points. But the amount of points must be complete). For example, ground truth have 10 key points, if student answer include one key he will get 0.5 point(if the answer include 5 points so should be 2.5). Medical student's answer: {answer} Key Points: {Key Point} Please only return a float number(from 0 to 5). You should check each point one by one(shouldn't judge based on language style such as fluence and so on. Only judge based on whether the student's answer include correct or relevant and complete key points). Don't generate any other information.

Table 7: Evaluation Pipeline Prompt Example Template.

Inference Score	Act as a USMLE evaluator, your role involves assessing and comparing a medical student's explanation to the provided target answer. Begin the assessment by carefully reviewing the provided target answer. Then, based on following specific criteria, determine the score for the student's answer. Please judge whether medical student's answer's diagnostic reasoning is correct based on ground truth. For example, ground truth have 10 steps, if student answer include one correct step he will get 0.5 point(if student have other correct diagnostic reasoning path it should also be correct. But the amount of evidence must be complete. It means that each step is about 0.5 point if there are 10 steps). Medical student's answer: {answer} Ground Truth: {diagnostic reasoning} Please only return a float number(from 0 to 5). You should check each step one by one(shouldn't judge based on language style such as fluence and so on. Only judge based on whether student's diagnostic reason is correct or relevant). Don't generate any other information.
Evidence Score	Act as a USMLE evaluator, your role involves assessing and comparing a medical student's explanation to the provided target answer. Begin the assessment by carefully reviewing the provided target answer. Then, based on following specific criteria, determine the score for the student's answer. Please judge whether medical student's answer provide detail evidence such as ground truth. For example, ground truth have 10 evidence, if student answer include one evidence he will get 0.5 point(if student give other correct detail evidence, it is also correct. But the amount of evidence must be complete.) Medical student's answer: {answer} Detail Evidence: {evidence} Please only return a float number(from 0 to 5). You should check each evidence one by one(shouldn't judge based on language style such as fluence and so on. Only judge based on whether student propose correct and complete diagnostic evidence). Don't generate any other information.

Table 8: Performance comparison across different models. Bold indicates best performance.

Model	Claude3-Opus	GPT-4o-mini	Huatuo-7B	Huatuo-34B	PMC-Cambrian	PMC-Cambrian-AN
Accuracy	58.4	46.2	34.5	44.7	41.2	**58.5**
UMLS Factuality	0.18	0.16	0.13	0.16	0.11	**0.23**
GPT-4 Overall	1.17	0.91	1.08	1.13	0.78	**1.29**
DeepSeek Overall	2.31	1.95	2.06	2.24	1.86	**2.55**

Dataset	**Data Size**	**Modality**	**ROI**	**Human Annotation**	**Slice ID**
Deeplesion	24,821	CT	×	×	×
PadChest	150,730	CT	×	✓	-
Eurorad	691,370	CT,X-Ray,MRI...(Multi)	✓	✓	✓
MIMIC-CXR-JPG	620,113	X-Ray	×	✓	-
LLD	30,390	MRI	✓	×	✓
MAMA-MIA	76,381	MRI	✓	×	✓
PMC-VQA	152,603	CT,X-Ray,MRI...(Multi)	×	✓	-
Path-VQA	19,654	Pathology	×	✓	-
PMC-Instruct	619,606	CT,X-Ray,MRI...(Multi)	×	✓	-
Quilt	1,017,416	Histopathology	×	✓	-
Radiopaedia	1,131,614	CT,X-Ray,MRI...(Multi)	✓	✓	✓
SLAKE	9,835	CT,X-Ray,MRI	×	✓	-
VQA-RAD	1,798	X-Ray,MRI	×	✓	-
AMBOSS & JAMA	45,820	Multi & Only Text	✓	✓	-
Chinese Data	300,000	Multi	-	-	-

Table 9: Data Source.

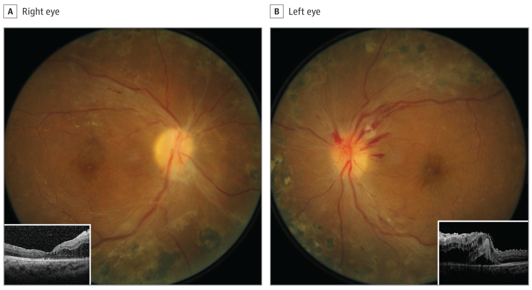
Table 10: Distribution of Articles Across JAMA Specialty Journals

Journal	**Count**
JAMA Otolaryngology–Head & Neck Surgery	513
JAMA Ophthalmology	466
JAMA Dermatology	368
JAMA (General)	328
JN Learning	299
JAMA Surgery	133
JAMA Oncology	105
JAMA Cardiology	92
JAMA Neurology	61
JAMA Pediatrics	60
JAMA Psychiatry	6

Dataset	**Caption Available**	**License**
DeepLesion	Yes	CC BY 4.0
PadChest	Yes	PADCHEST Dataset Research Use Agreement
Eurorad	Yes	Creative Commons Attribution 4.0 International License
MIMIC-CXR-JPG	No	PhysioNet Credentialed Health Data License 1.5.0
LLD	Yes	LLD-MMRI Agreement
MAMA-MIA	Yes	CC BY-NC-SA 4.0
PMC-VQA	Yes	CC BY-SA
PMC-Instruct	Yes	OpenRAIL
Quilt	Yes	-
Radiopaedia	No	Radiopaedia Agreement
JAMA Clinical Challenge	No	JAMA Agreement
LLaVA-Med	Yes	CC BY-NC 4.0

Table 11: Overview of caption availability and dataset licenses.

Table 12: Sample Case in JAMA Clinical Challenge.



Question: A woman in her mid-20s presented with subacute bilateral vision loss that was worse in the left eye. Her medical history was remarkable for type 1 diabetes diagnosed at 16 years of age and proliferative diabetic retinopathy in both eyes that had been treated with panretinal photocoagulation 7 years earlier. She had undergone pars plana vitrectomy with endolaser to treat a tractional retinal detachment in her right eye 2 years before this presentation. She also had a history of hypertension and chronic kidney disease, and she was 15 weeks into pregnancy. Visual acuity was 20/50 OD and 20/100 OS. Intraocular pressure was normal bilaterally, and no relative afferent pupillary defect was detected. Findings of an anterior segment examination were normal. The patient was in no apparent distress and denied any headache, chest pain, or focal weakness. Ophthalmoscopic examination (Figure) revealed mild optic nerve head edema that was greater in the left eye than the right eye with associated nerve fiber layer hemorrhage in the left eye. Nerve fiber layer infarctions, dot and blot hemorrhages, and lesions caused by panretinal photocoagulation also were seen bilaterally. Optical coherence tomography showed macular edema that involved the center of the macula in both eyes (Figure, inset). A. Obtain a fluorescein angiogram B. Determine blood glucose level and perform glycated hemoglobin test C. Measure heart rate, respiratory rate, and blood pressure D. Perform immediate computed tomography of the head Answer with the option's letter from the given choices directly and give me the reason. Answer with the option's letter from the given choices directly and give me the reason

Diagnostic Reason: Malignant hypertension with papillopathy C. Measure heart rate, respiratory rate, and blood pressure The patient was found to have hypertension, with a blood pressure of 195/110 mm Hg. Heart and respiratory rates were normal. Measurement of the arterial blood pressure may be performed rapidly in the clinic with a sphygmomanometer and is essential to rule out malignant hypertension, which is a potentially life-threatening cause of vision loss. Although the differential diagnosis for bilateral optic nerve edema is broad, workup should always include assessment of blood pressure when appropriate, because a hypertensive emergency (also known as malignant hypertension) may cause substantial morbidity or mortality if not diagnosed and treated promptly. Findings may include macular star, macular edema, serous retinal detachment, intraretinal hemorrhage, and optic disc edema with or without associated hemorrhage.[1,2] Optic nerve head edema may occur with systolic blood pressures as low as 160 mm Hg, with the median onset occurring at 188 mm Hg.[3] The macular edema associated with hypertensive retinopathy may be distributed more nasally, as was seen in this patient.[4] This patient had mild optic nerve edema despite high systemic blood pressure and substantial macular edema. This less-pronounced optic nerve edema likely was attributable to optic nerve atrophy at baseline. Proliferative diabetic retinopathy and panretinal photocoagulation can be associated with optic atrophy, and atrophic optic nerves tend to become less edematous than healthy optic nerves.[5-7] Regarding the other choices above, a fluorescein angiogram (choice A) would be expected to show leakage from the optic nerve and macula, but such findings are already available from the optical coherence tomography, which showed intraretinal and subretinal fluid. Although assessment of serologic levels of glucose and glycated hemoglobin (choice B) is important in the management of diabetic retinopathy, results from such tests are not immediately available and have less bearing on the immediate management of the case (this patient's glycated hemoglobin level was 10.5% [to convert to a proportion of total hemoglobin, multiply by 0.01]).

Performing computed tomography (choice D) would be a reasonable later step, especially if the patient's blood pressure was found to be normal, to rule out an intracranial mass. In addition to hypertension, other causes of optic nerve head edema exist. Papilledema (when optic nerve head edema is secondary to increased intracranial pressure) is often bilateral and may be caused by intracranial mass lesions, meningitis, cerebral venous thrombosis, or intracranial hypertension, or it may be idiopathic. Bilateral optic nerve head edema also may be caused by diabetic papillopathy (which is a rare cause of optic disc edema, but which is bilateral in up to 50% of cases[8,9]) and toxic optic neuropathies. Optic disc drusen are frequently bilateral and may cause so-called pseudopapilledema.[10] Severe hypertension resulting in acute retinopathy and optic nerve head edema is considered a hypertensive emergency because of its association with end-organ damage, namely, damage to the retina and optic nerve. This case serves as a reminder that, although the differential diagnosis for optic nerve head edema is broad, immediately assessing the patient's blood pressure to evaluate for hypertensive emergency in the eye clinic may be critical, because this treatable condition could lead to substantial morbidity or mortality if diagnosis is delayed. The patient was admitted to the medicine service for management of hypertension. At an ophthalmology follow-up examination 3 months later, her blood pressure was 128/73 mm Hg and her visual acuity was 20/40 OD and 20/60 OS, close to her visual acuity before this acute event. On examination, the optic nerve head edema had resolved, and optical coherence tomography showed resolution of macular edema.

PMC-Cambrian: Malignant hypertension with papillopathy C. Measure heart rate, respiratory rate, and blood pressure The key to the correct diagnosis is the presence of bilateral optic neuropathy in a pregnant woman. The patient's history of hypertension and chronic kidney disease, along with the findings of bilateral optic neuropathies, suggested a systemic etiology. A fluorescein angiogram (choice A) would not be the next step because the patient was asymptomatic and had no signs of retinal ischemia. A glycated hemoglobin test (choice B) would not be indicated because the patient had no signs of diabetic retinopathy. A computed tomography scan of the head (choice D) would not be appropriate because the patient had a normal examination and no signs of intracranial hemorrhage. Besides hypertension, other causes of optic nerve head edema exist. Papilledema, which is optic nerve head edema secondary to increased intracranial pressure, is often bilateral and may result from intracranial mass lesions, meningitis, cerebral venous thrombosis, intracranial hypertension, or it may be idiopathic. Bilateral optic nerve head edema can also be caused by diabetic papillopathy—a rare cause of optic disc edema that is bilateral in up to 50% of cases¹—and toxic optic neuropathies. Optic disc drusen, which are frequently bilateral, may lead to so-called pseudopapilledema. Severe hypertension causing acute retinopathy and optic nerve head edema is considered a hypertensive emergency due to its association with end-organ damage, specifically affecting the retina and optic nerve. This case emphasizes that, although the differential diagnosis for optic nerve head edema is extensive, promptly assessing the patient's blood pressure in the eye clinic to evaluate for a hypertensive emergency is critical. Immediate evaluation is crucial because this treatable condition could lead to significant morbidity or mortality if diagnosis is delayed. The patient was admitted to the medical service for hypertension management. At an ophthalmology follow-up examination three months later, her blood pressure was 128/73 mm Hg, and her visual acuity improved to 20/40 in the right eye (OD) and 20/60 in the left eye (OS), nearing her visual acuity before this acute event. Upon examination, the optic nerve head edema had resolved, and optical coherence tomography showed a resolution of the macular edema.

Claude3-Opus: This case presents several concerning features that make measuring vital signs the most urgent next step: Bilateral optic nerve edema (worse in left eye) in a young pregnant woman with:

- Type 1 diabetes
- Hypertension
- Chronic kidney disease
- Active pregnancy (15 weeks)

The key concern here is POSTERIOR REVERSIBLE ENCEPHALOPATHY SYNDROME (PRES), which is:

- A neurological condition characterized by vasogenic edema
- Commonly associated with severe hypertension
- More common in pregnant women
- Can present with visual symptoms and optic nerve edema
- Can be life-threatening if not treated promptly

Why other options are less appropriate:

- A (Fluorescein angiogram): Contraindicated in pregnancy
- B (Blood glucose/HbA1c): While important for diabetic management, not the most urgent concern given the presentation
- D (Head CT): While eventually needed if PRES is suspected, vital signs should be checked first as they'll guide immediate management

The vital signs are crucial because:

- They will help determine if this is a hypertensive emergency requiring immediate intervention.
- They're non-invasive and provide immediate critical information

The findings will guide the urgency and direction of further workup

- If severely elevated blood pressure is found, immediate treatment could prevent serious complications for both mother and fetus
- This represents a potential medical emergency where checking vital signs is the most appropriate first step to guide immediate management decisions.

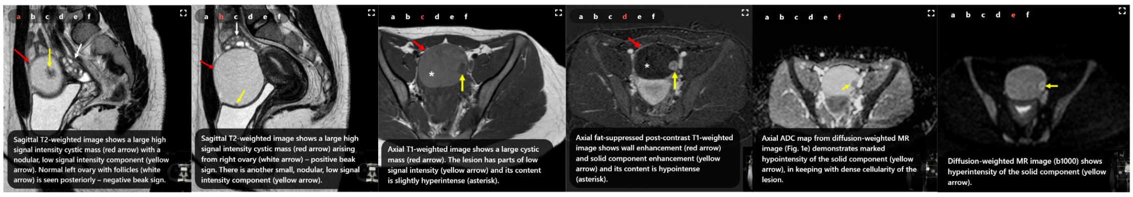
Huatuo-GPT-Vision-34B: The patient, who has a history of type 1 diabetes and is currently 15 weeks pregnant, presents with optic nerve head edema, macular edema, and hemorrhages in both eyes; therefore, it is crucial to assess her blood glucose levels and perform a glycated hemoglobin (HbA1c) test to evaluate and manage her glycemic control status, as this evaluation is essential to identify any acute changes or complications related to her pregnancy and existing diabetes, since poor glycemic control can lead to rapid progression of diabetic retinopathy, exacerbate ocular conditions, and pose significant risks to both maternal and fetal health—including the potential for congenital anomalies, preeclampsia, and worsening of chronic kidney disease—thus, timely assessment and adjustment of her diabetes management plan are vital to prevent further vision loss and ensure the well-being of both the mother and the developing fetus.

Huatuo-GPT-Vision-7B:B. Determine blood glucose level and perform glycated hemoglobin test Reasoning: The patient has a history of type 1 diabetes, which is known to cause diabetic retinopathy leading to vision loss as seen in the images provided (optic nerve head edema with hemorrhage). The presence of macular edema also suggests ongoing or poorly controlled diabetes affecting her eyes. A fluorescein angiogram would be useful for assessing vascular abnormalities but may not directly address the underlying issue related to poor glycemic control. Measuring heart rate, respiratory rate, and blood pressure are important general assessments but do not specifically relate to the ophthalmic findings presented here. Immediate computed tomography of the head might be considered if there was suspicion of other complications like intracranial pathology; however, it's less relevant given the current clinical context focused on eye examination results linked to diabetes management issues.

GPT-o-mini: Correct Answer: C. Measure heart rate, respiratory rate, and blood pressure. Reason: Measuring blood pressure is critical in this scenario to identify malignant hypertension, a potentially life-threatening condition that can lead to vision loss through optic nerve and retinal damage. Immediate assessment of blood pressure helps rule out hypertensive emergencies, allowing for timely intervention.

Expert Comment:PMC-Cambrian and Claude3-Opus accurately recognized the urgency of measuring vital signs to assess for malignant hypertension in this patient, who presents with bilateral optic nerve edema and a history of hypertension and chronic kidney disease during pregnancy. Their reasoning reflects a proper understanding of the need for immediate intervention to prevent serious complications. On the other hand, Huatuo-GPT-Vision-34B and Huatuo-GPT-Vision-7B focused on assessing glycemic control by selecting to determine blood glucose levels and perform a glycated hemoglobin test. While managing diabetes is important, they failed to prioritize the immediate life-threatening condition suggested by the patient's symptoms, thus overlooking the critical need to rule out a hypertensive emergency. GPT-4o-mini could get the correct answer and some key points but lack lots of detail evidence to prove it.

Table 13: Human Annotated Sample Case.



Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) with a nodular, low signal intensity component (yellow arrow). Normal left ovary with follicles (white arrow) is seen posteriorly – negative beak sign.

Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) arising from right ovary (white arrow) – positive beak sign. There is another small, nodular, low signal intensity component (yellow arrow).

Axial T1-weighted image shows a large cystic mass (red arrow). The lesion has parts of low signal intensity (yellow arrow) and its content is slightly hyperintense (asterisk).

Axial fat-suppressed post-contrast T1-weighted image shows wall enhancement (red arrow) and solid component enhancement (yellow arrow) and its content is hypointense (asterisk).

Axial ADC map from diffusion-weighted MR image (Fig. 1e) demonstrates marked hypointensity of the solid component (yellow arrow), in keeping with dense cellularity of the lesion.

Diffusion-weighted MR image (b1000) shows hyperintensity of the solid component (yellow arrow).

Image Caption: 1. Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) with a nodular, low signal intensity component (yellow arrow). Normal left ovary with follicles (white arrow) is seen posteriorly – negative beak sign. 2. Sagittal T2-weighted image shows a large high signal intensity cystic mass (red arrow) arising from right ovary (white arrow) – positive beak sign. There is another small, nodular, low signal intensity component (yellow arrow). 3. Axial T1-weighted image shows a large cystic mass (red arrow). The lesion has parts of low signal intensity (yellow arrow) and its content is slightly hyperintense (asterisk). 4. Axial fat-suppressed post-contrast T1-weighted image shows wall enhancement (red arrow) and solid component enhancement (yellow arrow) and its content is hypointense (asterisk). 5. Diffusion-weighted MR image (b1000) shows hyperintensity of the solid component (yellow arrow). 6. Axial ADC map from diffusion-weighted MR image (Fig. 1e) demonstrates marked hypointensity of the solid component (yellow arrow), in keeping with dense cellularity of the lesion.

Clinical History: A 21-year-old G0P0 woman with no medical history was referred to our institution for a sonographically detected cystic right adnexal mass. She has a history of pelvic discomfort without other complaints. Physical examination was normal. Laboratory findings were also normal except for an elevated CA 125 65.2 U/mL (normal <35.0).

Image Findings: MRI examination revealed a cystic tumour arising from the right ovary with 7.5 cm. On T2-weighted images, the signal intensity of the cyst content was high and two small nodular peripheral solid components were detected, adhering to its internal wall, with low signal (Fig. 1a, b). The normal left ovary was present with follicles (Fig. 1a). On pre-contrast T1-weighted images, the mass exhibited slightly high signal intensity (Fig. 1c). On contrast-enhanced fat-suppressed T1-weighted images, wall enhancement and solid component enhancement were detected (Fig. 1d). Finally, the ADC map (Fig. 1f) from diffusion-weighted image (Fig. 1e) demonstrates marked hypointensity of the solid component, in keeping with its dense cellularity. Surgical excision was proposed and accepted by the patient. The histopathological investigation revealed a typical ovarian serous borderline tumour.

Discussion: Borderline ovarian tumours are uncommon ovarian neoplasms, intermediate between benign and malignant types, corresponding to 5% of all epithelial ovarian tumours. [1, 2] Serous borderline tumour represents the most common type of borderline tumours arising in the ovary, and typically, it is confined to the adnexa and presents an indolent course. [3] However, up to 6.8% of these tumours can progress to low grade serous carcinoma. [3] Serous borderline tumours are divided into typical (90%) and borderline tumours with micro-papillary patterns (5%–10%). [4] These neoplasms usually present as bilateral adnexal masses with more proliferation of papillary projections than do benign cystadenomas, they are often seen in younger patients, and laboratory findings show the serum CA-125 level mildly elevated. [2, 3, 5, 6] The peak age of presentation is 45 years. [5] Small tumours usually do not cause symptoms and are often detected as an incidental finding on sonography. [7] Larger or more advanced neoplasms might cause pain or pelvic discomfort. The diagnosis of this type of tumour is based on histopathological examination. As they are staged using the same ovarian cancer staging of malignant ovarian neoplasms [5], MRI plays a crucial rule in this evaluation. There are no pathognomonic imaging features of borderline ovarian tumours and they frequently resemble a benign lesion. Nevertheless, some pathologic and MR imaging studies suggested that large papillary projections are highly suggestive of borderline or malignant neoplasms. [1, 8] The preferred treatment is bilateral oophorectomy. Fertility sparing procedures such as unilateral oophorectomy or cystectomy are related with a higher incidence of recurrence when compared with bilateral oophorectomy (10-20% vs. approximately 5% for radical surgery). [3, 4] Young age (<30 years) has also been found to be a risk factor for recurrent disease in patients treated conservatively. As in this case, patients should be instructed to complete child-bearing within a short time frame following initial diagnosis, due to the high risk of recurrence. [3] The most important feature in predicting an adverse outcome in serous borderline tumours is the presence of serous lesions involving the peritoneum (so-called invasive implants). [3, 4] Tumours without invasive implants, as in this case, have been considered to be nonaggressive tumours, associated with a good prognosis and high survival rate. Follow-up using ultrasound imaging is mandatory, with particular attention paid to the remaining ovary in conservatively treated young patients. [4] Written informed patient consent for publication has been obtained.

SemiHVision(only a sample for one image):

User: What type of imaging technique is shown in Fig. 1a of the provided image?

GPT: The image shows a T2-weighted MRI scan.

User: Where is the cystic tumour located according to the image?

GPT: The cystic tumour is arising from the right ovary

User: What is the size of the cystic tumour?

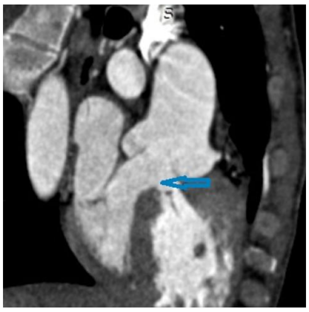
GPT: The cystic tumour is 7.5 cm in size.

User: What does the red arrow in the image indicate?

GPT: The red arrow indicates the normal left ovary with follicles.

User: What does the yellow arrow in the image point to?
User: The yellow arrow points to one of the small nodular peripheral solid components adhering to the internal wall of the cystic tumour.
User: What characteristic of the cyst content is shown by the high signal intensity in the T2-weighted image?
User: The high signal intensity indicates that the cyst content is fluid-filled.
User: What is shown by the asterisk (*) in the image?
User: The asterisk (*) indicates the cystic tumour arising from the right ovary.

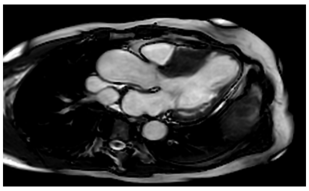
Table 14: Human Annotated Sample Case.

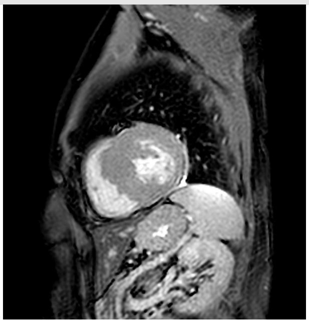


Prompt(without retriever): Generate the Image description based on the image.

GPT-4o: The image is a sagittal CT scan, likely of the thoracic or abdominal region. The scan shows various soft tissue structures, including major organs. The blue arrow points toward a specific structure, possibly a blood vessel or part of an organ. This could indicate an abnormality such as a mass, an aortic aneurysm, or an area of vessel narrowing or dissection. In the background, major anatomical landmarks like the heart, aorta, and vertebral column are visible, as well as some portions of the liver or spleen. The areas of high contrast indicate dense structures, such as bones, while softer tissues are seen in varying shades of gray, highlighting different densities. This scan could be used to evaluate conditions like vascular abnormalities, tumors, or other pathologies in the chest or upper abdominal region. Further clinical information would be necessary to confirm the exact diagnosis.

Prompt(with retriever): Generate the Image description based on the image. Here are some medical guidelines and Clinical cases for you.





Case Discussion: A case of asymmetric hypertrophic cardiomyopathy with mild subaortic stenosis and partial systolic anterior motion of the mitral valve. Since the pressure gradient obtained by phase contrast imaging was rather low, echocardiographic reevaluation and/or follow-up was recommended. The case also demonstrates a typical pattern of myocardial fibrosis in the basal and midventricular septal segments at the anterior and posterior right ventricular insertion points.

GPT-4o: The image is a coronal oblique CT scan of the thoracic region, focused on the heart and its surrounding structures. The blue arrow indicates a subaortic ventricular septal defect (VSD), measuring approximately 2.7 cm. This congenital heart defect refers to an abnormal opening in the interventricular septum (the wall separating the left and right ventricles of the heart), located just below the aortic valve. The surrounding anatomy includes parts of the heart chambers, likely showing the left and right ventricles, as well as nearby vascular structures such as the aorta. The defect appears to allow abnormal communication between the ventricles, which could lead to altered blood flow dynamics, potentially causing symptoms such as shortness of breath, fatigue, or heart failure if left untreated.