GMAI-VL & GMAI-VL-5.5M: A Large Vision-Language Model and a Comprehensive Multimodal Dataset Towards General Medical AI

Tianbin Li[1*], Yanzhou Su[1*], Wei Li[1,2], Bin Fu[1], Zhe Chen[1], Ziyan Huang[1,2], Guoan Wang[1], Chenglong Ma[1], Ying Chen[1], Ming Hu[1], Yanjun Li[1], Pengcheng Chen[1], Shixiang Tang[1], Xiaowei Hu[1], Zhongying Deng[1], Yuanfeng Ji[3], Jin Ye[1†], Yu Qiao[1], Junjun He[1†]

[1]Shanghai Artificial Intelligence Laboratory
[2]Shanghai Jiao Tong University
[3]Stanford University

{litianbin, yejin, hejunjun}@pjlab.org.cn

Abstract

Despite significant advancements in general AI, its effectiveness in the medical domain is limited by the lack of specialized medical knowledge. To address this, we formulate GMAI-VL-5.5M, a multimodal medical dataset created by converting hundreds of specialized medical datasets with various annotations into high-quality image-text pairs. This dataset offers comprehensive task coverage, diverse modalities, and rich image-text data. Building upon this dataset, we develop GMAI-VL, a 7B-parameter general medical vision-language model, with a three-stage training strategy that enhances the integration of visual and textual information. This approach significantly improves the model's ability to process multimodal data, supporting accurate diagnoses and clinical decision-making. Experiments show that GMAI-VL achieves state-of-the-art performance across various multimodal medical tasks, including visual question answering and medical image diagnosis.

Code — https://github.com/uni-medical/GMAI-VL

Datasets — https://huggingface.co/datasets/General-Medical-AI/GMAI-VL-5.5M

Introduction

Recent advancements in Large-scale Vision-Language Models (LVLMs) have driven progress in image recognition, natural language processing, and multimodal tasks, leveraging the power of multimodal datasets. In the medical field (general medical AI, GMAI), as these technologies mature, the need for accurate processing of diverse data—such as medical images, clinical text, and structured records—has become critical for reliable diagnostic and treatment decisions.

However, existing LVLMs, like GPT-4o (Achiam et al. 2023), exhibit limitations in medical applications due to their lack of domain-specific knowledge. This underscores the necessity for specialized approaches that integrate medical expertise. To bridge this gap, we construct a comprehensive medical vision-language dataset and develop a domain-specific model tailored for medical scenarios.

As shown in Fig. 1 (a), Our dataset provides high-quality medical knowledge across three aspects: (i) *Comprehensive medical tasks*: To improve the model's applicability in various medical scenarios, our dataset covers a wide range of disease types, symptoms, treatments, and medical workflows. (ii) *Rich multimodal representation*: Our dataset includes a variety of modalities, such as different types of medical images (e.g., CT, MRI, X-rays) and textual data (e.g., medical records, imaging reports). (iii) *High-quality image-text data*: The performance of medical LVLMs heavily relies on high-quality image-text pairs. We use the well-curated collection of medical images paired with precise textual descriptions to build our dataset.

To achieve this, we develop an annotation-guided data generation process to create this large-scale multimodal medical dataset. First, we collect multiple open-source medical imaging datasets and extract their key annotations (*e.g.*, modality, task type, labels, bounding boxes). Next, we convert these well-labeled annotations into text descriptions suitable for training vision-language models. This ensures that the medical image-text pairs are accurately constructed, with annotations verified by medical experts in the collected medical imaging datasets. The process results in the GMAI-VL-5.5M dataset, consisting of 5.5 million samples, which supports the development of general medical LVLMs.

Building upon this dataset, we develop a general medical vision-language model, GMAI-VL. To enhance its capacity for visual-linguistic integration and complex instruction following, we propose a three-stage training strategy. The first two stages focus on shallow and deep alignment, progressively aligning medical images and textual descriptions—from low-level visual features to high-level semantic representations. In the final stage, the model is fine-tuned using cross-modal instruction tuning, which significantly improves its understanding of visual-language interactions and its ability to perform complex, instruction-driven tasks, as

[*]Equal contribution
[†]Corresponding authors: {yejin, hejunjun}@pjlab.org.cn

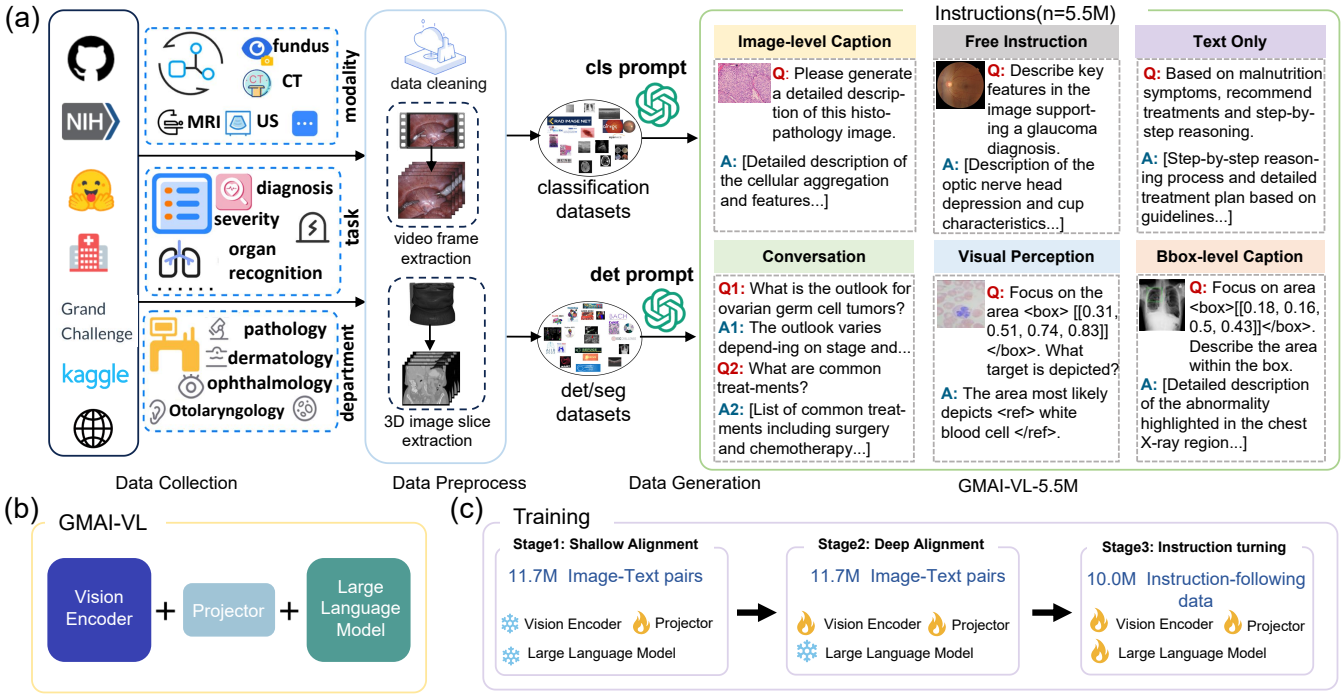


Figure 1: Overview of GMAI-VL and GMAI-VL-5.5M. (a) Sources, departments, modalities, task types, and instruction formats of GMAI-VL-5.5M. (b) Architecture of GMAI-VL, with a Vision Encoder, Projector, and Large Language Model. (c) Three-stage training process, including shallow alignment, deep alignment, and instruction tuning, with corresponding data sizes and components. 🔥 indicates the training part while ❄ indicates the frozen part.

illustrated in Fig. 1(b) and (c).

We conduct experiments on various multimodal medical benchmarks (Lau et al. 2018; He et al. 2020; Liu et al. 2021; Zhang et al. 2023; Chen et al. 2024b; Hu et al. 2024; Yue et al. 2024), and the results demonstrate that GMAI-VL significantly outperforms existing models. Specifically, GMAI-VL sets new benchmarks with an average score of 88.48% on OmniMedVQA, 62.43% on the GMAI-MMBench *test* set, and 51.3% on the Health & Medicine track of MMMU.

The contributions of this work are as follows:

- We develop an annotation-guided data generation methodology to create GMAI-VL-5.5M, a large-scale medical vision-language dataset derived from over 200 specialized medical datasets, covering diverse medical tasks, modalities, and high-quality image-text pairs.
- Leveraging this dataset, we design GMAI-VL, a general medical vision-language model, and propose a three-stage training strategy to enhance its integration of visual and linguistic features, improving its performance in various medical tasks.
- We demonstrate that GMAI-VL outperforms existing models in multimodal question-answering tasks, setting new benchmarks with high performance on OmniMed-VQA, GMAI-MMBench, and the Health & Medicine track of MMMU.

Related Work

Large-scale medical vision-language datasets. These datasets are crucial for training Large Vision-Language Models (LVLMs) in the medical domain. While general datasets are readily available, biomedical datasets often focus on text or images separately, limiting their generalization. Datasets like MIMIC-CXR (Johnson et al. 2019) and CheXpert (Chambon et al. 2024) advance radiology models but are restricted to a single image modality (X-ray), limiting their use as general-purpose medical LVLMs.

To address this, researchers have compiled large-scale vision-language datasets by scraping public resources like PubMed and medical textbooks. Datasets like LLaVA-Med (Li et al. 2024), Med-Flamingo (Moor et al. 2023), and PubMedVision (Chen et al. 2024a) improve upon LLaVA-Med by enhancing the quality of medical data. Additionally, open-source image datasets with annotations have been converted into image-text pairs for training. Notable examples include RadFM (Wu et al. 2023b), MedDr (He et al. 2024), BiomedGPT (Zhang et al. 2024), Med-Gemini (Saab et al. 2024), and Med-PaLM (Singhal et al. 2023). MedTrinity-25M (Xie et al. 2024) generates image-text pairs for supervised fine-tuning.

However, these datasets often face limitations in modalities, data sources, or task coverage, necessitating improvements. We construct a comprehensive medical vision-language dataset with broad task coverage, diverse modalities, and high-quality image-text pairs to provide a solid

foundation for model training.

Medical vision-language models. Medical vision-language models often adapt general-purpose Large Vision-Language Models (LVLMs) for specific medical tasks using specialized datasets. For example, Med-Flamingo (Moor et al. 2023) improves OpenFlamingo-9B with 0.8 million interleaved and 1.6 million paired medical image-text data for medical image analysis and report generation. RadFM (Wu et al. 2023b) enhances PMC-LLaMA (Wu et al. 2023a) with 16 million radiology images and text descriptions. Med-PaLM (Tu et al. 2024) adapts PaLM-E (Driess et al. 2023) to medical data, achieving state-of-the-art results in diagnostic support and Q&A. LLaVA-Med (Li et al. 2024) uses a biomedical figure-caption dataset from PubMed Central to improve LLaVA (Touvron et al. 2023a,b) for biomedical image understanding and open-ended conversations. Med-Gemini (Saab et al. 2024) leverages long-format question-answering datasets for better multimodal and contextual capabilities, enhancing complex medical Q&A and reasoning tasks. Additionally, HuatuoGPT-Vision (Chen et al. 2024a) and MedDr (He et al. 2024) adapt general-purpose LVLMs like LLaVA and InternVL to various medical modalities, including radiology, pathology, dermatology, and endoscopy.

While these studies focus on constructing medical datasets, they often overlook adaptation strategies. Naive training methods may fail to bridge the gap between natural and medical image-text pairs, or align diverse medical modalities and texts (e.g., prescriptions, radiology reports, EHRs), limiting generalizability. Our work introduces a novel three-stage training strategy to better integrate visual and language features, improving generalization.

GMAI-VL-5.5M: A Comprehensive Multimodal Medical Dataset

With rapid advancements in medical vision-language models, constructing high-quality datasets is essential for developing general-purpose models. Unlike previous approaches that mainly rely on published literature, our method leverages specialized medical datasets with various annotations to create a more robust, high-quality resource.

We present GMAI-VL-5.5M, a comprehensive medical vision-language dataset that aggregates data from diverse open-source and proprietary sources. Covering 13 medical imaging modalities and 18 specialties, it supports a wide range of medical imaging tasks. This dataset enhances the model's ability to process complex medical information, advancing precision medicine and intelligent diagnostics.

Data Curation

Data collection. To construct a comprehensive multimodal medical dataset, we sourced 219 datasets from diverse platforms. Fig. 1(a) highlights key data sources, such as Kaggle, Grand Challenge, and Huggingface, which facilitate extensive data collection. These datasets cover diverse imaging modalities, including fundus, CT, MRI, and ultrasound (US), and span a variety of medical tasks, such as diagnosis, severity assessment, and organ recognition. They also encompass multiple clinical specialties, including pathology, dermatology, ophthalmology, otolaryngology, and oncology, further enhancing their diversity.

Data processing. After data collection, we follow a workflow to pair medical data (including both 2D and 3D data) with corresponding text annotations. To ensure high-quality annotations, we first extract key information from the annotations provided by medical experts. For classification data, we extract the modality, department, and labels for each image, discarding instances with missing or unclear annotations. For segmentation data, we follow the SA-Med2D-20M (Ye et al. 2023a) approach, filtering out low-quality images and labels, and converting them into detection annotation format. Finally, the preprocessed data is standardized and organized into a structured format: `<image, modality, label, department, bbox [optional]>`, where "bbox" refers to the bounding box locations for detection annotations.

Paired data format generation. To generate paired visual medical data with text descriptions, we use large vision-language models (*i.e.*, GPT-4o) to produce detailed descriptions with instructions via an annotation-guided methodology. In details, for classification datasets, comprehensive descriptions of the entire image are created, while for detection datasets, the focus is on specific regions enclosed by bounding boxes, with detailed functional analyses of these areas. Furthermore, based on the given information, instruction-following question-answer pairs related to the medical images are generated. These pairs involve specific instructions tailored to the medical context, such as identifying critical anatomical or pathological features within the medical images (e.g., tumors, lesions, or organs) or providing in-depth interpretations of regions of interest. These instructions guide the model to produce relevant, precise responses, thereby enhancing its applicability to specialized medical imaging tasks. As mentioned in the previous subsection, the segmentation dataset is transformed into a detection format using external bounding boxes, and data generation follows the detection dataset protocols. The detailed example prompts are shown in Table 2. The generated data is then used for medical Visual Question Answering (VQA) tasks, forming the comprehensive VQA dataset, GMAI-VL-5.5M. To enhance the model's multilingual capability, we translate approximately 30% of the English image-text data into Chinese. This multilingual data helps further improve the generalization ability of domain-specific multimodal models.

Data Property

Data statistics. Our dataset encompasses a broad spectrum of medical imaging tasks and modalities, forming a robust foundation for the development and evaluation of medical LVLMs. Table 3 summarizes the distribution of key modalities, tasks, clinical departments, and clinical tasks within GMAI-VL-5.5M, illustrating its diversity and extensive coverage. *For a more detailed analysis, additional visual insights into the dataset composition can be found in the supplementary material.*

Data quality. The quality of the generated paired data and annotations is ensured through two main approaches.

Datasets	Data Size	Modality	Language	Traceability	Data Source
PathVQA (He et al. 2020)	32.7k	Pathology	EN	×	Textbooks
MIMIC-CXR (Johnson et al. 2019)	227k	X-Ray	EN	✓	Hospital
quilt-1M (Ikezogwo et al. 2024)	1M	Pathology	EN	×	YouTube & PubMed
MedDr VQA (He et al. 2024)	197k	Multimodal	EN	✓	13 medical datasets
PMC-OA (Lin et al. 2023)	1.65M	Multimodal	EN	×	PubMed
PMC-VQA (Zhang et al. 2023)	413k	Multimodal	EN	×	PubMed
LLaVA-Med VQA (Li et al. 2024)	56,702	Multimodal	EN	×	PubMed
ChiMed-VL (Liu et al. 2023b)	1.05M	Multimodal	CN	×	PubMed
PMC-CaseReport (Wu et al. 2023b)	438k	Multimodal	EN	×	PubMed
PubMedVision (Chen et al. 2024a)	1.29M	Multimodal	EN&CN	×	PubMed
GMAI-VL-5.5M (ours)	5.5M	Multimodal	EN&CN	✓	219 specialized medical imaging datasets

Table 1: Comparison of multimodal medical datasets, including size, modality, language, traceability, and sources.

First, we use high-quality datasets from trusted sources, including professional challenges like Kaggle and Grand-Challenge, as well as peer-reviewed datasets, ensuring data accuracy and reliability. Second, we carefully control the data generation process. While GPT is used for generation, the prompts are designed with essential annotations (e.g., `<image, modality, label, department, bbox [optional]>`) to minimize errors. This annotation-guided approach produces more detailed and professional descriptions. *The data quality is detailed in the supplementary material.*

Compared with other medical multimodal datasets. The GMAI-VL-5.5M dataset, as shown in Table 1, distinguishes itself with its unmatched scale, featuring over 5.5 million samples from more than 219 specialized medical imaging datasets. Unlike other datasets, GMAI-VL-5.5M supports a broader range of modalities and languages, making it a truly global resource that addresses diverse clinical needs. Furthermore, GMAI-VL-5.5M ensures data traceability, maintaining high clinical relevance and reliability. This extensive and varied dataset is crucial for advancing medical multimodal research, enabling more effective training of LVLMs that can generalize across numerous medical tasks and scenarios, ultimately driving innovations in precision medicine and intelligent diagnostics.

GMAI-VL: A General Medical Vision-Language Model

Architecture

The GMAI-VL model is a vision-language model built upon the LLaVA architecture (Liu et al. 2023a; Li et al. 2024), incorporating three key components: a large language model (LLM), a vision encoder, and a projector (MLP), as illustrated in Fig. 1(b).

We use InternLM2.5-7B (Team 2023) as our language processing module, which provides exceptional reasoning capabilities. With a context window of up to one million tokens, it can manage complex medical tasks and generate coherent, accurate responses. For vision processing, we adopt a CLIP-based vision encoder (Radford et al. 2021), which converts visual inputs into high-dimensional feature representations. The MLP projector bridges the vision encoder and language processing module, optimizing high-dimensional outputs and improving feature representation. This framework enables GMAI-VL to effectively process multimodal medical data.

Training Strategy

As shown in Fig.1(c), the training process of the GMAI-VL model is divided into three stages: shallow alignment, deep alignment, and instruction tuning. To enhance the training of GMAI-VL, we supplement the GMAI-VL-5.5M dataset with additional medical datasets, resulting in a total training dataset of 11.7M samples.

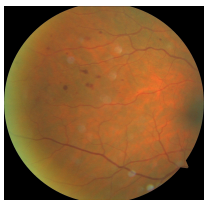
Stage I: Shadow alignment. In this phase, we use a large-scale medical image-text dataset of approximately 11.7 million image-text pairs. During this stage, we freeze both the large language model and the vision encoder, optimizing only the projector. This optimization establishes an initial alignment between medical images and their corresponding textual descriptions.

During training, (i) the objective is to minimize the cross-entropy loss of the text tokens; (ii) all images are resized to 336×336 pixels; (iii) the learning rate is set to $1e^{-3}$ with a cosine decay schedule, and AdamW is used as the optimizer; (iv) the total batch size is $32 \times 8 \times 2$, where 32 refers to the number of GPUs used, eight represents the micro-batch size per GPU, and two is the number of gradient accumulation steps; and (v) a soft packing technique is used to allow each sample to contain multiple sequences, averaging over two sequences per sample. We train this stage for 1 epoch.

Stage II: Deep alignment. Most vision encoders in multimodal models are pre-trained on natural images, so addressing the domain gap between medical and natural images is critical during the deep alignment stage. To bridge this gap, we fine-tune both the vision-language projector and the vision encoder, enhancing the alignment between the visual features of medical images and the language model's feature space. The learning rate is set to $1e^{-4}$, the total batch size is $32 \times 4 \times 4$, and other settings remain consistent with Stage I. We train this stage for 1 epoch.

Stage III: Instruction tuning. In this stage, we fine-tune the entire GMAI-VL model—vision encoder, language model, and projector—through instruction tuning to enhance its instruction-following and dialogue capabilities. The multimodal instruction data is primarily derived from the previous stages. Note that this stage discards the low-quality samples that with too short descriptions or extremely

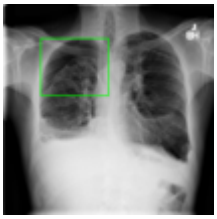
Instruction-following data prompt



As a medical expert, you will receive a <fundus>image belonging to <Ophthalmology>, which is labeled as <moderate nonproliferative diabetic retinopathy>.
Please generate question-answer pairs based on the following requirements:
1. The question-answer pairs must be strictly based on the content of the image and directly related to the label information, should not state known medical facts or general definitions directly.
2. Focus on reasoning around the label and guide observation of the visible features of the disease, such as affected areas, symptoms, and changes in the retina.
3. Each question and answer must be precise, clear, and derived from observable characteristics in the image, with no vague or uncertain information.
4. The output format must follow the standard JSON format with the structure as follows:
{ 'question': <question content >, 'answer': <answer content >}

Question *Can we see any hard exudates or fluid leakage in the image?*
Answer *The image does show some bright, yellowish deposits, which are consistent with hard exudates, often seen in moderate nonproliferative diabetic retinopathy due to fluid leakage.*
Question *What is the appearance of the optic disc in this retinal image?*
Answer *The optic disc appears normal without any visible signs of swelling or abnormal changes, suggesting that the optic nerve is not yet affected by the disease.*

Caption data prompt



As a medical expert, you are given a <X-ray> image belonging to <Pulmonary Medicine>, where the <bbox [179,164,500,434]> is labeled as <Pneumothorax>.
Please generate a very detailed description related to bbox in the given image. When generating the description, you must strictly adhere to the following guidelines:
1. The generated description should be inferable from the image;
2. The generated description should include image features related to the given label;
3. The generated description should start with 'This region';
4. The description may imply known information but should not directly state known information;
5. The generated description should include major entities within and around the region, background elements, texture features, and any other notable features.

Output: This region shows *... the right hemithorax on a frontal chest X-ray*, *... an area suggestive of abnormal air accumulation*. *The bounded area appears less dense compared to surrounding lung tissue*, *... The edges of the lung are visible, separated from the chest wall by a faint radiolucent (darker) space, suggestive of pneumothorax. The lung markings are notably absent within this area*, *... The texture within the defined area has a homogeneous, darkened appearance consistent with air rather than tissue density, contrasting with the granular pattern in the aerated lung parenchyma outside this region.*

Table 2: The examples of annotation-guided prompts for paired data format generation.

Dimension	Value	Percentage
Task Type	2D Classification	50.4%
	3D Segmentation	30.3%
	2D Segmentation	12.7%
	2D Detection	6.6%
Modality	CT	26.8%
	MR	24.7%
	Endoscopy	12.6%
	Pathology	11.2%
	X-Ray	6.7%
	Fundus	5.3%
	Ultrasound	3.0%
Department	Orthopedic Surgery	12.9%
	General Surgery	10.3%
	Gastroenterology	9.7%
	Hematology	9.2%
	Pulmonary Medicine	9.0%
	Sports Medicine	8.2%
Clinical Task	Disease Diagnosis	40.4%
	Organ Recognition	16.0%
	Bone Recognition	8.5%
	Severity Grading	6.1%
	Surgeon Action Recognition	6.0%

Table 3: Distribution of GMAI-VL-5.5M across key dimensions.

Model	VQA-RAD	SLAKE	PMC-VQA	Avg.
Med-Flamingo (Moor et al. 2023)	45.4	43.5	23.3	37.4
RadFM (Wu et al. 2023b)	50.6	34.6	25.9	37.0
LLAVA-Med-7B (Li et al. 2024)	51.4	48.6	24.7	41.6
Qwen-VL-Chat (Bai et al. 2023)	47.0	56.0	36.6	46.5
Yi-VL-34B (Young et al. 2024)	53.0	58.9	39.5	50.5
LLAVA-v1.6-7B (Liu et al. 2024)	52.6	57.9	35.5	48.7
LLAVA-v1.6-13B (Liu et al. 2024)	55.8	58.9	36.6	50.8
LLAVA-v1.6-34B (Liu et al. 2024)	58.6	67.3	44.4	56.8
HuatuoGPT-V-7B (Chen et al. 2024a)	<u>63.8</u>	**74.5**	<u>52.7</u>	**<u>63.7</u>**
GMAI-VL (w/o our data)	62.3	66.3	39.0	55.9
GMAI-VL (ours)	**66.3**	<u>72.9</u>	**54.3**	**64.5**

Table 4: Results on medical VQA benchmarks. The highest performance in each column is highlighted in **bold**, and the second-highest performance is highlighted in <u>underlined</u>.

lower confidence in the previous stages. Additionally, medical text dialogue data is incorporated to improve the model's handling of various dialogue scenarios. This results in ten million samples for instruction tuning. During taining, the learning rate is set to $1e^{-5}$, the total batch size is $32 \times 4 \times 4$, while other parameters, including optimizer, remain unchanged across stages. We train this stage for 2 epochs.

Experimental Results

To evaluate our model, we utilized several established multimodal medical benchmarks, including medical VQA benchmarks (PMCVQA (Zhang et al. 2023), PathVQA (He et al.

2020), VQA-RAD (Lau et al. 2018), and SLAKE (Liu et al. 2021)) as well as benchmarks designed for large vision-language models (OmniMedVQA (Hu et al. 2024), GMAI-MMBench (Chen et al. 2024b), and the MMMU Health & Medicine track (Yue et al. 2024)). These benchmarks target specific aspects of medical image understanding and question answering. *Detailed information about these benchmarks can be found in the supplementary material.*

We evaluate model performance using VLMEvalKit (Duan, Yang et al. 2024) with its default settings. To prevent data leakage in GMAI-VL-5.5M, we ensure that the evaluation datasets, OmniMedVQA and GMAI-MMBench, are sourced exclusively from public test sets, while the training data is strictly from public training sets. Additionally, MD5 hashes were computed for each image to verify that there are no duplicates with the benchmark images.

Comparisons on Medical VQA Datasets

The performance of various VLMs on popular medical VQA benchmark datasets is summarized in Table 4. Our model, GMAI-VL (ours), demonstrates strong performance, achieving the highest score of 66.3% on the VQA-RAD (Lau et al. 2018) dataset, outperforming models such as HuatuoGPT-Vision-7B. This result underscores GMAI-VL's superior capability in handling radiological image question-answering tasks. On the PMC-VQA (Zhang et al. 2023) dataset, GMAI-VL achieves 54.3%, and 72.9% on SLAKE (Liu et al. 2021), highlighting its effectiveness across diverse medical VQA tasks. Overall, GMAI-VL demonstrates competitive performance across multiple benchmarks, showcasing its versatility in medical image understanding and question answering.

Comparisons on OmniMedVQA

The OmniMedVQA (Hu et al. 2024) benchmark integrates 73 traditional medical imaging datasets, all formatted for visual question answering (VQA). Table 5 summarizes the performance of various large vision-language models (LVLMs), including GMAI-VL, across five question types: Modality Recognition, Anatomy Identification, Disease Diagnosis, Lesion Grading, and Other Biological Attributes.

From the results, we can see that (i) GMAI-VL excels in all tasks, achieving 98.64% in Modality Recognition, 92.95% in Anatomy Identification, and 88.71% in Disease Diagnosis; (ii) it outperforms both open-source and medical-specific models, demonstrating strong abilities in identifying anatomical structures and diagnosing diseases; (iii) in Lesion Grading, it achieves the highest score of 87.21%, and scores 82.95% in Other Biological Attributes, showcasing its versatility; and (iv) with an average accuracy of 88.48%, GMAI-VL surpasses models like HuatuoGPT-Vision-34B and InternVL2-40B, establishing itself as a leading model in multimodal medical image understanding and setting a new benchmark for medical VQA tasks.

Comparisons on GMAI-MMBench

The GMAI-MMBench benchmark is a comprehensive evaluation suite for medical multimodal models, focusing on clinical visual question-answering (VQA) tasks. Table 6 presents the performance of various models on the *val* and *test* sets across a range of clinical tasks. Notably, (i) GMAI-VL excels in tasks such as abnormality recognition (73.78%), biological variation recognition (63.06%), and clinical disease diagnosis (66.67%), demonstrating its strong ability to understand and interpret complex clinical images. (ii) Compared to other models, GMAI-VL consistently ranks first or second across most tasks, securing the top position in 16 out of 20 categories. (iii) Key tasks such as Attribute Recognition (AR) and Disease Diagnosis (DD) yield scores of 75.26% and 67.14%, respectively, highlighting GMAI-VL's strength in medical scenario understanding. (iv) Overall, GMAI-VL sets a new benchmark in various clinical VQA tasks.

Model	MR	AI	DD	LG	OBA	Overall
Random Guess	25.00	25.84	28.41	25.40	37.49	28.28
Open-Source LVLMs						
MiniGPT-4 (Zhu et al. 2023)	36.98	32.68	24.19	20.45	26.14	27.59
LLaVA (Liu et al. 2023a)	52.30	35.27	11.80	9.77	24.70	22.86
LLaMA_Adapter_v2 (Gao et al. 2023)	58.45	38.18	29.12	23.73	30.97	35.08
InstructBLIP (Dai et al. 2024)	72.35	39.90	32.01	43.80	47.91	41.14
BLIP-2 (Li et al. 2023)	57.48	49.83	46.21	30.52	73.52	50.77
Qwen-VL-Chat (Bai et al. 2023)	33.69	10.95	16.27	6.71	41.68	20.29
mPLUG-Owl2 (Ye et al. 2023b)	78.01	48.52	39.68	20.56	59.36	48.44
LLaVa-NeXT (Liu et al. 2024)	68.23	46.74	41.21	18.43	39.57	45.57
DeepSeek-VL (Lu et al. 2024)	74.01	51.94	45.46	21.06	29.04	48.76
Yi-VL (Young et al. 2024)	59.56	44.81	48.97	32.93	24.63	47.28
InternVL2-40B (Chen et al. 2024c)	96.76	64.25	76.28	76.50	76.27	78.70
Medical Special Model						
MedVInT-TE (Zhang et al. 2023)	62.62	41.03	40.57	12.17	45.17	43.83
LLaVA-Med (Li et al. 2024)	48.41	27.96	23.72	16.10	21.94	27.82
Med-Flamingo (Moor et al. 2023)	26.74	25.10	23.80	28.04	16.26	23.82
RadFM (Wu et al. 2023b)	27.45	21.65	23.75	16.94	20.05	23.48
MedDr (He et al. 2024)	91.37	51.62	65.56	73.18	74.52	68.27
HuatuoGPT-V-34B (Chen et al. 2024a)	95.06	75.67	66.51	72.83	74.92	73.23
medgemma-4b (Sellergren et al. 2025)	95.25	84.01	81.26	83.08	60.61	81.92
Our Model						
GMAI-VL (w/o our data)	96.40	80.97	79.14	70.29	75.66	79.96
GMAI-VL (ours)	**98.64**	**92.95**	**88.7**	**87.21**	**82.95**	**88.48**

Table 5: Comparison of LVLMs and GMAI-VL on OmniMedVQA across five question types. The best performance in each column is highlighted in **bold**, and the second-best in underlined. **Abbreviations**: MR = Modality Recognition, AI = Anatomy Identification, DD = Disease Diagnosis, LG = Lesion Grading, OBA = Other Biological Attributes.

Comparisons on MMMU Health & Medicine

We further assess the performance of our GMAI-VL on the Health & Medicine track of MMMU benchmark, which is a widely recognized standard for evaluating multimodal models. The experimental results in Table 8 show the model's performance across five key categories: Basic Medical Science (**BMS**), Clinical Medicine (**CM**), Diagnostics and Laboratory Medicine (**DLM**), Pharmacy (**P**), and Public Health (**PH**).

The results show that (i) GMAI-VL performs strongly across multiple categories, achieving top scores in **DLM** (43.3%), **P** (50.0%), and **PH** (53.3%), surpassing competitive models such as LLaVA-v1.6 and HuatuoGPT-Vision-7B. These results highlight the model's proficiency in handling complex tasks that require diagnostic reasoning, pharmaceutical knowledge, and public health expertise. (ii) In **BMS**, GMAI-VL scores 50.0%, achieving the best perfor-

Model Name	Overall (val)	Overall (test)	AR	BVR	B	CR	C	DD	IQG	MR	M	NT	OR-A	OR-HN	OR-P	OR-T	SG	SAR	SIR	SWR
Medical Special Model																				
Med-Flamingo (Moor et al. 2023)	12.74	11.64	6.67	10.14	9.23	11.27	6.62	13.43	12.15	6.38	8.00	18.18	9.26	18.27	11.00	11.53	12.16	5.19	8.47	11.43
LLaVA-Med (Li et al. 2024)	20.54	19.60	24.51	17.83	17.08	19.86	15.04	19.81	20.24	21.51	13.20	15.15	20.42	23.73	17.67	19.65	21.70	19.81	14.11	20.86
Qilin-Med-VL-Chat (Liu et al. 2023b)	22.34	22.06	29.57	19.41	16.46	23.79	15.79	24.19	21.86	16.62	7.20	13.64	24.00	14.67	12.67	15.53	26.13	24.42	17.37	25.71
RadFM (Wu et al. 2023b)	22.95	22.93	27.16	20.63	13.23	19.14	20.45	24.51	23.48	22.85	15.60	16.16	14.32	24.93	17.33	21.53	29.73	17.12	19.59	31.14
MedDr (He et al. 2024)	41.95	43.69	41.20	50.70	37.85	29.87	28.27	52.53	36.03	31.45	29.60	47.47	33.37	51.33	32.67	44.47	35.14	25.19	25.58	32.29
HuatouGPT-V-7B (Chen et al. 2024a)	50.20	51.03	53.04	58.27	47.47	39.93	37.44	58.77	42.00	45.99	29.20	69.00	51.93	57.16	53.07	55.65	33.61	26.73	27.76	43.43
medgemma-4b (Sellergren et al. 2025)	36.20	35.49	38.96	38.18	29.54	31.91	16.69	38.38	30.00	37.24	26.80	42.00	36.89	39.35	33.60	38.59	35.59	20.96	24.93	38.29
Our Model																				
GMAI-VL (w/o our data)	54.99	56.23	51.26	**61.05**	53.79	44.39	44.51	62.60	40.80	57.42	35.20	**79.50**	**61.31**	**77.81**	**53.60**	**69.29**	35.39	35.77	29.71	44.86
GMAI-VL (ours)	**61.74**	**62.43**	**75.26**	59.66	**67.24**	**56.86**	**54.29**	**67.14**	**42.80**	**79.97**	**41.60**	75.00	60.45	75.48	53.33	58.12	**42.09**	**72.31**	**37.40**	**59.14**

Table 6: Results on the *val* and *test* sets of GMAI-MMBench for clinical VQA tasks. Full task names are in Table 5 of (Chen et al. 2024b), and additional model comparisons are included in the Supplementary Materials. The best and second-best models are marked in **bold** and underlined, respectively.

Model	MMMU_val	OmniMedVQA	GMAI_MMBench_val	GMAI_MMBench_test
stage1	46.00	81.38	51.49	53.69
stage1+2	46.67	84.64	55.85	58.28
stage1+3	45.33	86.62	59.25	60.70
stage1+2+3 (Prop)	**51.30**	**88.48**	**61.74**	**62.43**

Table 7: Evaluation on the training strategy.

Model	BMS	CM	DLM	P	PH	MMMU (H&M)
Med-Flamingo	33.6	30.2	23.3	29.3	25.8	28.4
RadFM	31.6	28.6	26.7	26.2	26.8	27.9
LLaVA-Med-7B	33.8	32.3	26.7	40.7	43.3	38.6
Qwen-VL-Chat	32.7	20.6	19.3	29.6	33.3	31.7
Yi-VL-34B	48.1	55.6	36.7	35.4	31.3	48.2
LLaVA-v1.6-7B	46.4	43.4	30.0	29.6	26.7	33.1
LLaVA-v1.6-13B	**53.6**	46.7	33.3	22.2	40.0	39.3
HuatuoGPT-V-7B	50.0	**63.3**	36.7	48.1	**53.3**	50.3
MedGemma-4b	43.3	50.0	40.0	36.7	46.7	43.3
GMAI-VL (Base)	43.3	56.7	**43.3**	46.7	40.0	46.0
GMAI-VL (Ours)	50.0	60.0	**43.3**	**50.0**	**53.3**	**51.3**

Table 8: Performance on the *val* set for the MMMU Health & Medicine track. The best performance is highlighted in **bold** while the second-best performance is highlighted in underlined. Note that the results are obtained from the official website.

mance and demonstrating its ability to understand medical knowledge. (iii) In **CM**, the model scores 60.0%, remaining competitive with other leading models. These results underscore the model's effectiveness in processing both clinical and foundational medical information. (iv) Overall, GMAI-VL achieves an average score of 51.3% across the Health & Medicine track, ranking among the top models and confirming its versatility in specialized medical domains.

Ablation Study of the GMAI-VL-5.5M Dataset

In this section, we evaluate the effectiveness of the proposed GMAI-VL-5.5M dataset. We report the results of GMAI-VL (w/o our data) in Tables 4, 5, 6, and 8. This model is trained using data excluding the GMAI-VL-5.5M dataset (see Sec.), highlighting that our dataset effectively enhances the performance of large vision-language models.

The results demonstrate that the GMAI-VL-5.5M dataset provides highly accurate and reliable medical knowledge, especially in recognizing and understanding multimodal medical data. This significantly improves model performance, showcasing the dataset's diversity, comprehensiveness, and its ability to complement other datasets in complex medical tasks.

Ablation Study of Training Strategy

Table 7 presents results at different training stages. (i) The proposed stage1+2+3 (Prop) training strategy significantly enhances model performance across multiple benchmarks, outperforming other models and training strategies in all key metrics. (ii) By progressively incorporating stages 1, 2, and 3, the model shows consistent improvements on various datasets. This highlights the effectiveness of the three-stage training pipeline in improving the model's ability to handle complex multimodal medical tasks, confirming that a more comprehensive training approach yields superior results.

Conclusion

In this paper, we build GMAI-VL, a 7B-parameter vision-language model, and GMAI-VL-5.5M, a comprehensive multimodal medical dataset designed to advance general medical AI. GMAI-VL-5.5M converts hundreds of medical image analysis datasets into high-quality image-text pairs through annotation-guided data generation, enabling GMAI-VL to tackle a wide range of clinical tasks effectively. Experimental results show that GMAI-VL-5.5M significantly enhances GMAI-VL's performance, achieving state-of-the-art results across multiple key benchmark datasets. Future work will focus on expanding the dataset with more diverse and challenging medical scenarios, further improving the model's ability to generalize across different clinical environments and applications.

Acknowledgments

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Chambon, P.; Delbrouck, J.-B.; Sounack, T.; Huang, S.-C.; Chen, Z.; Varma, M.; Truong, S. Q.; Chuong, C. T.; and Langlotz, C. P. 2024. CheXpert Plus: Hundreds of Thousands of Aligned Radiology Texts, Images and Patients. *arXiv preprint arXiv:2405.19538*.

Chen, J.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; Yu, G.; et al. 2024a. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. *arXiv preprint arXiv:2406.19280*.

Chen, P.; Ye, J.; Wang, G.; Li, Y.; Deng, Z.; Li, W.; Li, T.; Duan, H.; Huang, Z.; Su, Y.; et al. 2024b. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. *arXiv preprint arXiv:2408.03361*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024c. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Duan, H.; Yang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

He, S.; Nie, Y.; Chen, Z.; Cai, Z.; Wang, H.; Yang, S.; and Chen, H. 2024. MedDr: Diagnosis-Guided Bootstrapping for Large-Scale Medical Vision-Language Learning. *arXiv preprint arXiv:2404.15127*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Ikezogwo, W.; Seyfioglu, S.; Ghezloo, F.; Geva, D.; Sheikh Mohammed, F.; Anand, P. K.; Krishna, R.; and Shapiro, L. 2024. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

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

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 525–536. Springer.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning.

Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; and Hua, Y. 2023b. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Team, I. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *NEJM AI*, 1(3): AIoa2300138.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023a. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2(5): 6.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023b. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Xie, Y.; Zhou, C.; Gao, L.; Wu, J.; Li, X.; Zhou, H.-Y.; Liu, S.; Xing, L.; Zou, J.; Xie, C.; et al. 2024. MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine. *arXiv preprint arXiv:2408.02900*.

Ye, J.; Cheng, J.; Chen, J.; Deng, Z.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023a. SA-Med2D-20M Dataset: Segment Anything in 2D Medical Imaging with 20 Million masks. *arXiv preprint arXiv:2311.11969*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 1–13.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.