Towards a Multimodal Large Language Model with Pixel-Level Insight for Biomedicine

Xiaoshuang Huang[*1, 2], **Lingdong Shen**[3], **Jia Liu**[1], **Fangxin Shang**[1], **Hongxiang Li**[4], **Haifeng Huang**[1], **Yehui Yang**[1 †]

[1] Baidu Inc
[2] China Agricultural University
[3] Institute of Automation, Chinese Academy of Sciences
[4] Peking University
huangxs497@gmail.com

Abstract

In recent years, Multimodal Large Language Models (MLLM) have achieved notable advancements, demonstrating the feasibility of developing an intelligent biomedical assistant. However, current biomedical MLLMs predominantly focus on image-level understanding and restrict interactions to textual commands, thus limiting their capability boundaries and the flexibility of usage. In this paper, we introduce a novel end-to-end multimodal large language model for the biomedical domain, named MedPLIB, which possesses pixel-level understanding. Excitingly, it supports visual question answering (VQA), arbitrary pixel-level prompts (points, bounding boxes, and free-form shapes), and pixel-level grounding. We propose a novel Mixture-of-Experts (MoE) multi-stage training strategy, which divides MoE into separate training phases for a visual-language expert model and a pixel-grounding expert model, followed by fine-tuning using MoE. This strategy effectively coordinates multitask learning while maintaining the computational cost at inference equivalent to that of a single expert model. To advance the research of biomedical MLLMs, we introduce the Medical Complex Vision Question Answering Dataset (MeCoVQA), which comprises an array of 8 modalities for complex medical imaging question answering and image region understanding. Experimental results indicate that MedPLIB has achieved state-of-the-art outcomes across multiple medical visual language tasks. More importantly, in zero-shot evaluations for the pixel grounding task, MedPLIB leads the best small and large models by margins of 19.7 and 15.6 respectively on the mDice metric.

Code — https://github.com/ShawnHuang497/MedPLIB

Introduction

Owing to their impressive capabilities in image understanding and text generation, models such as GPT-4V and LLaVA (Liu et al. 2024) within the realm of Multimodal Large Language Models (MLLMs) have garnered widespread research interest from both academic and industrial sectors (Chen et al. 2023; You et al. 2023). Numerous researchers have dedicated efforts to explore the potential applications of MLLMs in the biomedical field (Wu et al. 2023a; Zhang et al. 2023; Moor et al. 2023), including LLaVA-Med (Li et al. 2024) and Med-PaLM M (Tu et al. 2024). MLLMs not only generate high-quality responses but also analyze biomedical imagery, demonstrating significant potential to transform traditional medical paradigms (Li et al. 2024; He et al. 2024). For doctors, such chatbots could significantly alleviate their heavy workloads and enhance efficiency. For patients, it provides more convenient access to professional medical knowledge and advice (Liu et al. 2023a). Additionally, this could also help alleviate the uneven distribution of medical resources, particularly in regions where they are scarce.

Unlike the image-level VQA of MLLMs in the natural world, the medical domain requires a finer-grained pixel-level understanding to ensure accuracy and answer interpretability. However, existing medical MLLMs (Li et al. 2024; Wu et al. 2023a; Moor et al. 2023; Tu et al. 2024) capacity remains restricted to image-level understanding, falling short of pixel-level perceptions. Pixel-level MLLMs offer several advantages over image-level perception: **Firstly**, they are capable of recognizing and processing detailed information by focusing on each pixel within an image, such as small lesions or subtle changes in tissue structures. **Secondly**, they facilitate improved structural recognition and pixel grounding. Pixel-level perception provides more precise grounding outcomes, aiding physicians in better understanding pathological images and formulating treatment plans. **Thirdly**, pixel-level analysis enables models to understand contextual information on a finer scale. Overall, pixel-level perception is particularly vital in multimodal large models within the medical field, especially in applications requiring high accuracy and detailed recognition.

Pixel-level MLLMs in biomedicine currently face significant challenges due to these two issues: **(1) Data scarcity:** Owing to privacy regulations and the high cost of labeling, there is a severe scarcity of pixel-level and complex VQA data. Openly available VQA datasets are typically designed for image-level multiple-choice questions or sim-

[*]Work performed during an internship at Baidu Inc.
[†]Corresponding author.

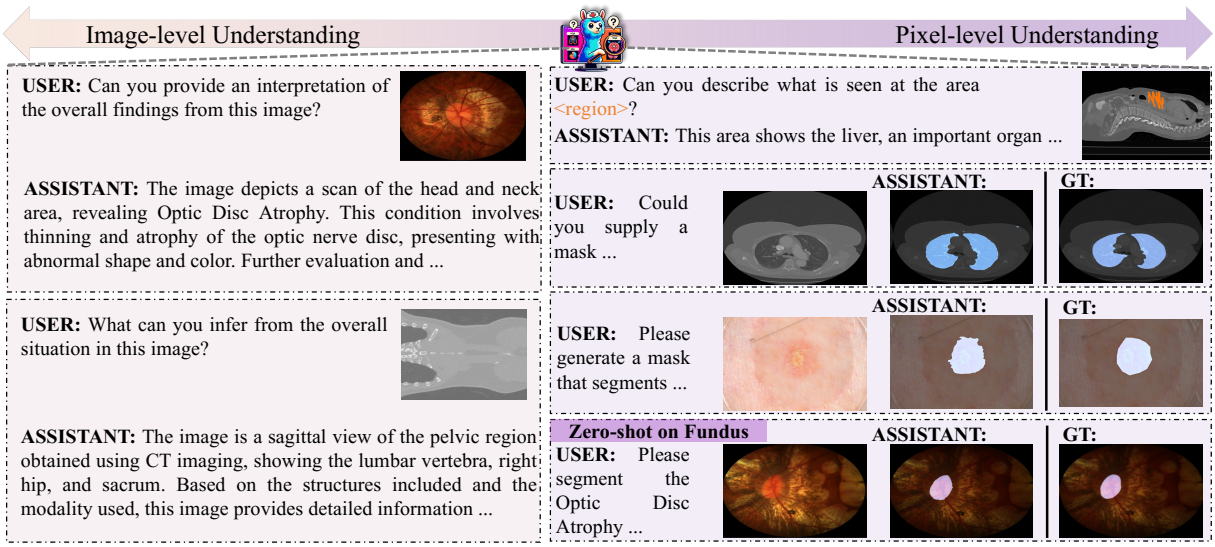

Figure 1: MedPLIB is a biomedical MLLM with a huge breadth of abilities and supports multiple imaging modalities.

ple question-and-answer formats, lacking potential for pixel analysis. Meanwhile, segmentation datasets usually contain only segmentation masks and simple category labels, devoid of textual semantic information. **(2) Models:** Medical VQA often requires a combination of spatial understanding (pixel-level understanding) to ensure confidence and interoperability. Integrating both knowledge-based question-answering and pixel-level analysis within the same MLLM is extremely challenging. Such multimodal input and output not only demand high architectural flexibility from the model but also require the model to balance the knowledge and capabilities of different tasks within a constrained parameter space.

For the first challenge, we propose MeCoVQA dataset. it amasses a substantial collection of segmentation datasets with category labels and, uses a Large Language Model (LLM) in conjunction with manual processing. Specifically, we first convert all segmentation masks of the images into structured metadata, which includes modality, body part, image orientation, and category labels corresponding to mask instances. Then, we use this metadata as a prompt to provide the LLM, which generates a comprehensive description of the image. Finally, we integrate the metadata and image description back as a prompt to the LLM to generate complex question-and-answer data. **For the second challenge**, in terms of framework, we expand the LLM's vocabulary and incorporate a region projector to extend the input modalities of the MLLMs. Additionally, inspired by LISA (Lai et al. 2024), we introduce a "<SEG>" token to identify and extract features necessary for pixel grounding, combined with SAM-Med2D (Cheng et al. 2023) to expand the response modalities of the MLLMs. It is noted that the overall framework is end-to-end. To better accommodate tasks of varying granularity, such as pixel grounding tasks and visual question-answering tasks, we introduce a novel multi-granular MoE training strategy within MLLMs that incorporates expert prior knowledge. Specifically, we train two experts separately for VQA and pixel grounding tasks. Subsequently, we integrate the two distinct experts via a training router.

In summary, our contributions are as follows: **(1) Model.** We propose an end-to-end MLLM with pixel-level insight for biomedicine, named MedPLIB. It simultaneously supports VQA, pixel-level prompts (points, bounding boxes, and free-form shapes), and pixel-level grounding. Experimental results indicate that MedPLIB achieves state-of-the-art outcomes across multiple medical visual language datasets. **(2) MeCoVQA Dataset.** It comprises an array of 8 modalities with a total of 310k pairs for complex medical imaging question answering and image region understanding. **(3) Open-source.** The data, codes, and model checkpoints will be released to the research community.

Related Work

Biomedical Visual Question Answering. Medical VQA can be categorized into two types based on the scale of the model and data size. Early approaches employ Convolutional Neural Networks (Simonyan and Zisserman 2014), long short-term memory models (Hochreiter and Schmidhuber 1997) followed by feature fusion method (Kim, Jun, and Zhang 2018) to generate response. Additionally, the transformer-based approaches like BERT (Devlin et al. 2018) and BioBert (Lee et al. 2020) achieve impressive performance. However, Due to the scarcity of medical data and model size, these models were prone to overfitting, leading to suboptimal robustness (Liu et al. 2023b). The emergence and success of MLLMs across broad applications have been well explored in the natural world (Liu et al. 2024; Chen et al. 2023; You et al. 2023). Parallel to these developments, the biomedical community has been zealously advancing the capabilities of MLLMs. A focal point of recent research has been the specialized domain of MLLMs, where significant advances have been made, particularly highlighted by mod-

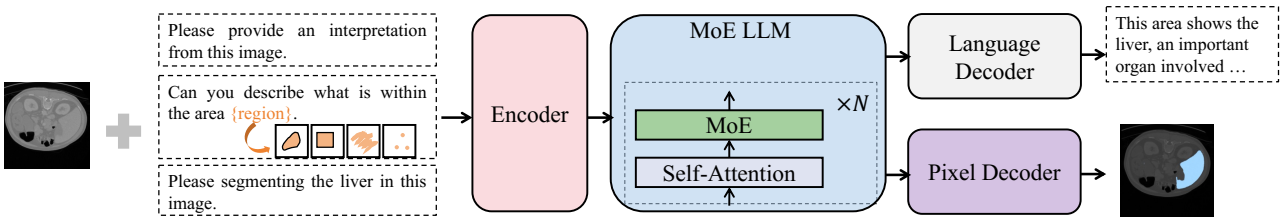

Figure 2: Overview of the proposed MedPLIB. It consists of three parts: encoder, MoE LLM, and decoder.

els such as RadFM (Wu et al. 2023a), LLaVA-Med (Li et al. 2024), and others (Luo et al. 2023; Liu et al. 2023a). These innovations have significantly advanced the biomedical applications of MLLMs. Although these models show strong task transfer capabilities, they are limited to fixed image and text inputs and outputs. This paper proposes expanding MLLMs to handle diverse input modalities (images, text, free-shape region prompts) and outputs (text, masks) to fully harness their potential.

Biomedical Image Segmentation. Over the decades, medical image segmentation has evolved significantly, with specialist small models under specific imaging modalities, such as U-Net (Ronneberger, Fischer, and Brox 2015), TransUnet (Chen et al. 2021), and Swin-Unet (Cao et al. 2022), achieving commendable results. However, the robustness and generalizability of these specialist models are suboptimal, restricting their application across multiple medical imaging modalities simultaneously. Recent research has shifted focus towards generic medical image segmentation (Zhang and Liu 2023; Wu et al. 2023b) and text-guided pixel grounding (Li et al. 2023b; Huang et al. 2024), yet both are hindered by data scarcity and model capacity, limiting performance enhancements. Unlike these approaches, our work is pioneering in the medical image analysis domain as we explore the expansion of pixel grounding capabilities within MLLMs, demonstrating the potential to circumvent the limitations of previous methods.

Mixture of Experts. MoE models have been proposed to augment the number of model parameters without incurring additional computational costs in machine learning (Jacobs et al. 1991; Fedus, Zoph, and Shazeer 2022). A series of approaches naturally decouple experts based on modal categories and pre-define each expert to handle a specific task (Long et al. 2023). This allows for the efficient utilization of shared model parameters. However, it necessitates manual switching between the required experts and lacks adaptive coordination among tasks. Recently, researchers in the fields of multimodal and natural language processing have focused on the study of soft routers (Chen et al. 2024). Soft MoE systems could lead the model to adaptively adjust between experts based on the input data and achieve model sparsity. The work most relevant to our architecture includes MoE-LLaVA (Lin et al. 2024) and LLMBind (Zhu et al. 2024), where all experts possess and are limited to the same prior knowledge. our paper focuses on resolving conflicts among various tasks. We advocate for distinct experts to possess independent task-specific prior knowledge and to coordinate among different tasks effectively.

Method

Architecture

The overall framework of MedPLIB comprises three structural layers: the encoder, the MoE LLM, and the decoder, as illustrated in Figure 2.

Encoder The encoder aims to encode all types of inputs (image, text, visual prompt) into a unified feature space.

Vision Tower and Pixel Encoder. Given the input image $v \in \mathbb{R}^{H\times W\times 3}$, we utilize the pre-trained CLIP visual encoder CLIP-ViT-L/336 (Radford et al. 2021) with a vision-projector as vision tower to extract the feature $V \in \mathbb{R}^{C_v\times N_v}$, where $N_v = \frac{H}{14} \times \frac{W}{14}$ and C_v is the hidden size of vision tower. Then project it as $\hat{V} \in \mathbb{R}^{C_{llm}\times N_v}$, where C_{llm} is the hidden size in the MoE LLM. Similarly, we employ a pre-trained Visual Transformer (ViT) (Dosovitskiy et al. 2020) with medical adapter layers (Cheng et al. 2023) as pixel encoder to get the pixel features $V_p \in \mathbb{R}^{C_p\times N_v}$, where C_p denotes the hidden size in the pixel decoder.

Vision Prompt Encoder. This block aims to appropriately prompt the LLM with user-specified areas of interest (boxes, points, free shapes) as inputs. Inspired by SEEM (Zou et al. 2024), we define a vision sampler to convert all types of non-textual queries into visual prompts that reside within the same visual embedding space. Assuming the area of interest input is R and m is the sampling pixel number, the visual prompt features V_{vp} can be formatted as $V_{vp} = MLP(\phi(V, m))$, where ϕ and MLP is the random sampling function and linear function.

Text Prompt Embedding. Inspire by LLaVA (Liu et al. 2024), we expand the tokenizer's vocabulary with "<region>" and "</region>" tokens to better integrate visual and textual prompts while distinguishing between their types. Assuming the input text is processed through the text prompt embedding layer to yield textual features $T \in \mathbb{R}^{C_{llm}\times N_t}$, where N_t is the embedded text length. We then use the embeddings of "<region>" and "</region>" to encapsulate the visual prompts V_{vp}. These are then embedded into specified positions in the textual feature sequence to obtain $\hat{T} \in \mathbb{R}^{C_{llm}\times \hat{N}_t}$, where $\hat{N}_t = N_t + 1$.

Finally, we concatenate $\hat{V}$ and $\hat{T}$ to obtain the output $X \in \mathbb{R}^{C_{llm}\times L}$ of the encoding module, where $L = \hat{N}_t + \hat{N}_v$.

Large Language Model with MoE For the given input X, the operation of the plain feed-forward layer with LoRA

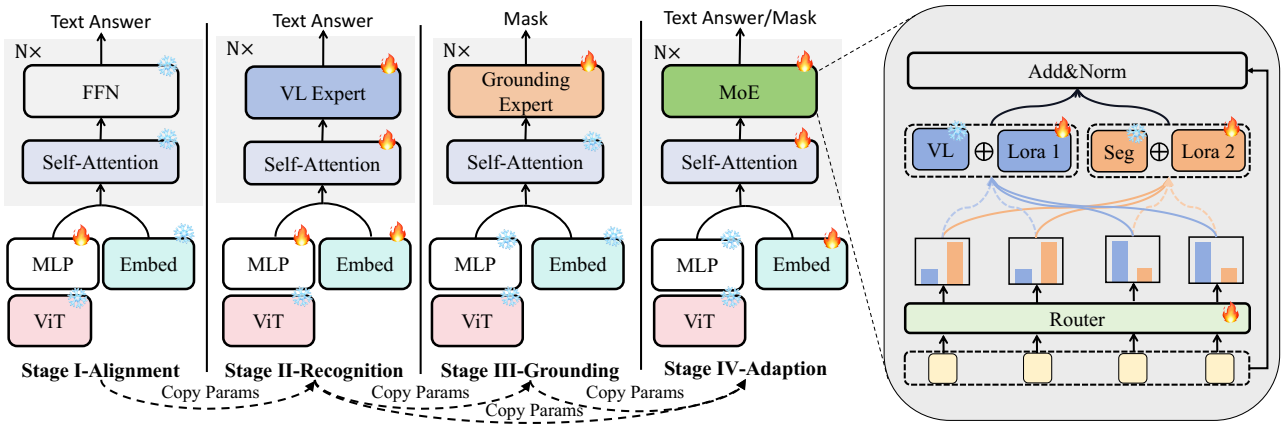

Figure 3: The multi-granular training strategy and the MoE block.

can be abstracted as $\hat{X} = WX + \Delta WX$, where $W \in \mathbb{R}^{C_{llm} \times C_{llm}}$ is the fixed parameter and $\Delta W \in \mathbb{R}^{C_{llm} \times C_{llm}}$ denotes the parameter update in the training phase.

To better accommodate tasks of varying granularity, such as pixel grounding tasks and visual question-answering tasks, we have introduced the MoE into the Feed Forward Network (FFN) within the LLM. Let's consider the vision-language expert as E_{vl} and the grounding expert as E_{ground}. The forward process of MoE layer can be formulated as:

$$\hat{X} = G(X)(\hat{X}_{vl} + \hat{X}_{ground}) \tag{1}$$

$$G(X) = Softmax(W_g \cdot X) \tag{2}$$

$$\hat{X}_{vl} = G(X)(E_{vl}X + \frac{\alpha}{r}\Delta W_0 X) \tag{3}$$

$$\hat{X}_{ground} = G(X)(E_{ground}X + \frac{\alpha}{r}\Delta W_1 X) \tag{4}$$

where $G(\cdot)$ denotes the router network in the MoE layer and W_g is a trainable parameter. The α and r is hyper-parameter. The $W_0 = B_0A_0$ and $W_1 = B_1A_1$ where $A_0, A_1 \in \mathbb{R}^{C_{llm} \times r}$, $B_0, B_1 \in \mathbb{R}^{r \times C_{llm}}$. Therefore, each token in X is processed by the top-1 expert with the highest probability, and the weighted sum is calculated based on the probabilities of the router.

Decoder For text, we use a linear layer as the Decoder, similar to common language models. For pixel-level grounding decoding, we follow LISA (Lai et al. 2024). We extract the last-layer embedding $\hat{h}_{ground}$ corresponding to the "<SEG>" token and utilize the T-projector to obtain h_{ground}. Finally, we use the SAM-Med mask decoder γ to obtain the prediction mask M. This process can be formulated as $M = \gamma(h_{ground}, V_p)$.

Multi-stage Training As shown in Figure 3, we present the multi-stage training strategy.

Stage I-Alignment: Following LLaVA-Med (Li et al. 2024) and LLaVA (Liu et al. 2024), we consider only the cross-entropy loss $\mathcal{L}_{reg}$ for text responses during this stage.

Stage II-Recognition: We train the MLLM as a base model to create an MLLM proficient in medical knowledge and medical imagery understanding. In this stage, we tackle complex visual-language tasks, such as visual knowledge multiple-choice questions, intricate medical Q&A, and region-based visual question answering. Specifically, we use vast question-and-answer pairs to fine-tune all modules except the vision tower. Thus, we have obtained a MLLM enriched with extensive medical imaging knowledge, where the FFN can be designated as E_{vl}. Similar to stage I, the training objective is to minimize the loss $\mathcal{L}_{reg}$.

Stage III-Grounding: To enhance the pixel grounding capability of the model, we focus on training the grounding expert in this stage. We use the model obtained from stage II as the initial model. We then train using the MeCoVQA-G dataset specifically targeting the FFN layer, pixel decoder, and T-projector. Ultimately, we achieve an MLLM equipped with pixel grounding knowledge, where the FFN is designated as E_{ground}. In this stage, we use binary cross-entropy $\mathcal{L}_{bce}$ and dice loss $\mathcal{L}_{dice}$ for pixel grounding losses, and $\mathcal{L}_{reg}$ for the text responses associated with the "<SEG>" token.

Stage IV-Adaption: After completing stage II and stage III, we obtain the parameters for E_{vl}, E_{ground}, and other modules. We then mix all available data and unfreeze all parameters, employing LoRA for fine-tuning through expert mixing. Using the router $G(\cdot)$, tokens are distributed to different experts for collaborative processing. This approach not only maintains minimal computational expenditure but also preserves the distinct prior knowledge of each expert. During mixed training, the optimization objective can be formulated as:

$$\mathcal{L} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice} \tag{5}$$

where λ_{reg}, λ_{bce}, $\mathcal{L}_{dice}$ are the hyperparameters to balance different objectives.

MeCoVQA Dataset

Large models are increasingly used to generate high-quality data, addressing data scarcity. However, in medical imaging, open-source datasets for detailed question-and-answer interactions remain limited. These are vital for intelligent biomedical assistants who need to perform detailed medical analyses and interact with patients. We suggest a new strategy for creating such detailed interactive data. MoCoVQA

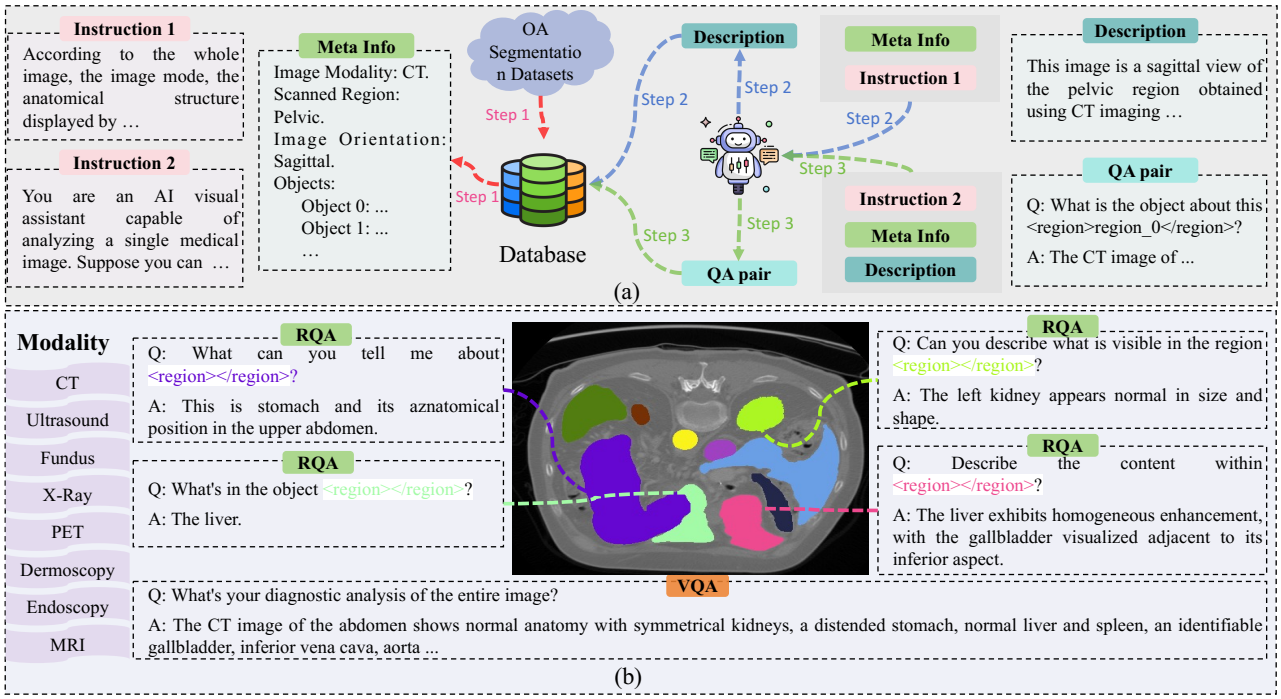

Figure 4: The construction pipeline (a) and a sample (b) of the MoCoVQA dataset.

was generated through the collaborative efforts of humans and an AI assistant, derived from large-scale biomedical image segmentation datasets. As shown in Figure 4, the generation process can be divided into three steps:

I. Manually generating instance-level meta information for each image based on its mask. We randomly sampled 100k biomedical images with instance masks from the SA-Med2D-20M (Ye et al. 2023a). Then we enrich the images with additional details to compile the meta information, which includes modality, scanned region, orientation, and object instances.

II. We use an AI assistant to get global descriptions for images, adjusting prompts to produce 500 data points per modality, which are manually reviewed for quality. We finalize the prompts only when all points meet quality standards.

III. Utilizing the AI assistant to craft pixel-level conversations based on the meta information and global descriptions obtained in step II. At this step, we used complex instructions to generate diverse data, manually refining prompts multiple times, as in stage II, to ensure quality.

The MeCoVQA dataset could be divided into three subsets: MeCoVQA-C (MeCoVQA-Complex), MeCoVQA-R (MeCoVQA-Region), and MeCoVQA-G (MeCoVQA-Grounding), which are used for complex VQA, region VQA, and pixel grounding, respectively. Complex VQA and region VQA are constructed through the aforementioned pipeline. MeCoVQA-G is generated by specifying question templates combined with mask category labels. Overall, the training set numbers for MeCoVQA-C, MeCoVQA-R, and MeCoVQA-G are 80k, 126k, and 100k respectively. Additionally, the numbers of their corresponding test sets are 1477, 2633, and 2344, respectively. For more information about MeCoVQA datasets, please refer to the Appendix.

Experiments

Experimental Setup

Model Settings. We employ SAM-Med2D (Cheng et al. 2023) as the pixel encoder and mask decoder. We use LLaMA-7B (Touvron et al. 2023) as a base LLM. Following LLaVA 1.5 (Liu et al. 2024), we utilize CLIP-Large (Radford et al. 2021) as the vision tower and the MLP consists of two linear layers with GELU activation function (Hendrycks and Gimpel 2016). The parameters of the model with 2 experts are 12 Billion. The training durations for stages I to IV are 9, 17, 15, and 77 hours, respectively. We provide additional model details in the Appendix.

Datasets. For the training data in stage I, we employ LLaVA-Med-alignment (Li et al. 2024). We utilize the union of MeCoVQA-R, MeCoVQA-C, SLAKE (Liu et al. 2021), PathVQA (He 2021), PMC-VQA (Zhang et al. 2023), ImageClef2021 (Ben Abacha et al. 2021), ImageClef2019 (Abacha et al. 2019), and VQA-RAD (Lau et al. 2018) in stage II. In stage III, we use MeCoVQA-G for training. Finally, we employ all data used in stages II and III for training in stage IV. The data volumes for stages I to IV are 330k, 400k, 100k, and 500k, respectively.

Metrics. For closed-set VQA, we report accuracy. For open-set VQA, we report precision and recall. For pixel-level grounding, we report the mean dice score.

Model	Param.	OmniMedVQA Benchmark									MeCoVQA Test			
		CT	MR	OCT	Der	MIC	X-Ray	FP	US	Mean	MeCoVQA-C		MeCoVQA-R	
MiniGPT-4 (Zhu et al. 2023)	7B	23.67	28.65	33.62	41.28	29.31	37.15	42.46	26.42	32.82	-	-	-	-
BLIP-2 (Li et al. 2023a)	4B	59.90	43.47	69.57	40.93	51.46	**64.97**	**67.61**	39.05	54.62	-	-	-	-
InstructBLIP (Dai et al. 2024)	7B	29.48	36.13	45.54	**63.02**	48.65	58.14	44.32	**43.35**	46.08	-	-	-	-
LLaVA (Liu et al. 2024)	7B	18.33	28.87	37.21	49.67	28.7	28.35	34.05	23.16	31.04	30.39	41.00	18.01	44.98
VPGTrans (Zhang et al. 2024)	7B	22.88	26.47	27.23	45.42	25.53	44.18	36.83	27.49	32.00	-	-	-	-
RadFM (Wu et al. 2023a)	14B	27.93	24.71	33.96	38.32	26.27	26.60	31.41	16.54	28.22	-	-	-	-
LLaVA-Med (Li et al. 2024)	7B	19.55	30.49	38.96	46.42	29.27	32.41	43.13	30.37	33.83	19.88	33.94	15.47	34.86
LISA (Lai et al. 2024)†	7B	62.96	49.01	66.19	41.61	54.70	62.34	46.71	32.10	51.95	56.63	52.83	12.31	13.20
MedPLIB-w/o MoE	7B	**63.22**	50.12	67.24	43.37	54.98	62.87	49.55	33.70	53.13	56.66	**52.94**	54.60	52.87
MedPLIB	12B/7B	62.70	**66.97**	**75.05**	51.47	**64.40**	60.25	65.04	38.75	**60.58**	**58.49**	49.41	**64.92**	**63.84**

Table 1: Performance on VQA. For closed-set OmniMedVQA (Hu et al. 2024), we report accuracy metrics. For open-ended MeCoVQA, we report precision (left) and recall (right) metrics. "CT", "MR", "OCT", "Der", "Mic", "US", "FP" denote Computed Tomography, Magnetic Resonance Imaging, Optical Coherence Tomography, Dermoscopy, Microscopy Images, Fundus Photography, and Ultrasound, respectively. The "A/B" format in the column of Parameters(Param.) indicates activated parameters during training and inference. † represents that we implement by office open-source code on our MeCoVQA dataset.

Model	Param.	MeCoVQA-G Test				Zero-shot					
		Der	CT	PET	Mean	X-Ray	End	MR	US	FP	Mean
LViT (Li et al. 2023b) TMI'23	30M	84.37	58.10	74.45	72.31	23.15	11.87	12.18	0.46	15.13	12.56
ReclMIS (Huang et al. 2024) Arxiv'24	74M/24M	88.81	74.96	**81.15**	81.64	26.62	10.19	3.54	0.00	11.29	10.33
LAVT (Yang et al. 2022) CVPR'22	119M	92.59	77.34	79.13	83.02	17.51	4.09	1.04	0.00	0.12	4.55
DMMI (Hu et al. 2023) ICCV'23	115	**93.38**	**79.97**	80.63	**84.66**	18.46	0.68	1.97	0.00	0.04	4.23
LISA (Lai et al. 2024) CVPR'24	7B	81.33	52.68	54.20	62.74	17.02	32.42	11.99	14.37	7.64	16.69
MedPLIB-w/o MoE	7B	79.66	55.92	56.97	64.18	20.92	36.43	15.87	17.98	11.31	20.50
MedPLIB	14B/7B	79.90	59.83	64.59	68.11	**28.25**	**44.19**	**27.52**	**35.64**	**25.76**	**32.27**

Table 2: Performance on MeCoVQA-G test set and zero-shot on cross modalities. The MeCoVQA-G test set compromises three modalities (CT, dermoscopy, PET). "End" denotes Endoscopy. The "A/B" format in the column of Parameters(Param.) indicates activated parameters during training and inference. All results in this table are implemented by office open-source code on our MeCoVQA-G dataset.

Performance Evaluation

Performance on VQA Benchmark. OmniMedVQA (Hu et al. 2024) is a large medical VQA benchmark that utilizes single-choice questions. We present the evaluation results of the open-source portion of the OmniMedVQA benchmark in Table 1. Across seven modalities, MedPLIB leads the second-best model, BLIP-2 (Li et al. 2023a), by an advantage of 7.84 points in the mean performance. Additionally, our MedPLIB significantly outperforms other biomedical MLLMs. For more analysis of this table, please refer to the Appendix.

Complex VQA Evaluation. Compared to test sets like OmniMedVQA (Hu et al. 2024), MeCoVQA-C features longer open-ended questions. As indicated in the third and fourth columns at the end of Table 1, our MedPLIB achieves better precision compared to LISA (Lai et al. 2024), but slightly lower recall. We believe this is due to MedPLIB balancing its pixel grounding capabilities, which slightly compromises its ability to handle long-text VQA tasks.

Region-level VQA Evaluation. Region-level VQA demands pixel-level image understanding. Current models lack support for region-specific prompts. To enable this, we integrate coordinates into the prompts for models like LLaVA (Liu et al. 2024), LLaVA-Med (Li et al. 2024), and LISA (Lai et al. 2024). As demonstrated in the last two columns of Table 1, our MedPLIB significantly outperforms these models.

Pixel Grounding Evaluation. Since there has not yet been a biomedical MLLM with pixel grounding capabilities, we compare our MedPLIB with small models that possess pixel grounding capabilities and with the influential LISA (Lai et al. 2024) from the general domain. As shown in the second column of Table 2, our MedPLIB surpasses LISA (Lai et al. 2024) by 5.37 points on the mDice metric in the MeCoVQA-G. However, it performs closely to the small model LVIT (Li et al. 2023b) and significantly lags behind DMMI (Hu et al. 2023).

Zero-shot to Pixel Grounding. To evaluate the generalization capabilities of our model, we conducted zero-shot assessments on five medical imaging modalities that the model did not see. As shown in the last six columns of Table 2, our MedPLIB demonstrated remarkable generalization capabilities. It significantly outperformed the best model on the MeCoVQA-G Test set, LViT (Li et al. 2023b), and pixel-grounding MLLM (LISA (Lai et al. 2024)). This underscores the substantial potential of our approach in addressing the generalization challenges that small models for medical image grounding struggle to overcome.

	stage I	stage II	stage III	stage IV	MeCoVQA-C	MeCoVQA-R	OmniMedVQA	MeCoVQA-G	Mean
(a)	-	-	-	-	60.18	58.88	53.30	34.11	51.62
(b)				✓	58.01	62.91	57.55	38.48	54.24
(c)	✓			✓	58.76	**63.74**	**57.95**	37.22	54.42
(d)	✓	✓		✓	58.25	63.24	57.00	39.63	54.53
(e)	✓		✓	✓	58.20	57.91	56.00	**47.52**	54.91
(f)	✓	✓	✓	✓	**60.55**	62.14	57.90	46.73	**56.83**

Table 3: Effect of the different training stages.

CF	Top-k	MeCo-VQA-C	MeCo-VQA-R	OmniM-edVQA	MeCo-VQA-S	Mean
1	1	57.95	59.23	55.85	44.92	54.49
1.5	1	**60.55**	**62.14**	**57.9**	**46.73**	**56.83**
2	1	55.35	55.94	56.15	46.64	53.52
2	2	55.05	57.64	54.35	45.32	53.09

Table 4: Effect of the hyper-parameters. "CF" denotes the capability factor of the expert. "Top-k" represents distributing the token to the top k experts with the highest probabilities for processing.

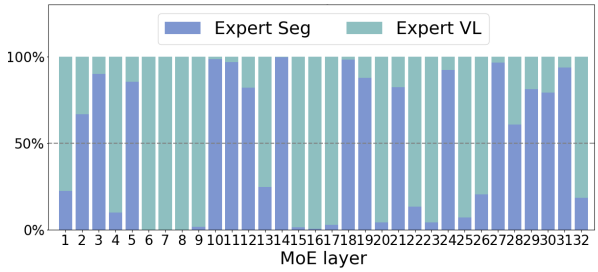

Figure 5: Distribution of tokens among different experts.

Qualitative Results

Figure 1 illustrates the performance of MedPLIB across various capabilities, addressing many issues beyond the scope of existing biomedical MLLMs. Additionally, we visualized the distribution of tokens among different experts in Figure 5. Overall, each expert processed about half of the tokens. This indicates that in MedPLIB, E_{vl} and E_{ground} achieved a good level of collaboration and load balancing. Additionally, we provide more results in the Appendix.

Ablation Study

We investigated the impact of MoE, training stages, and key hyperparameters on model performance in this section. It is important to note that conducting ablation experiments on all data is prohibitively costly, so during the ablation stage, we used a training set consisting of 20k samples extracted from the total dataset. For testing, we extracted 400 samples from the original MeCoVQA-C and MeCoVQA-R test sets and 2000 samples from OmniMedVQA (Hu et al. 2024).

Effect of MoE. The variants (a) and (b) in Table 3 display the performance of using a standard FFN and a MoE, respectively. Overall, the average performance of MoE across four datasets is 2.62 points higher than that of FFN, demonstrating MoE's adaptability to our tasks.

Effect of Multi-stage Training. In Table 3, we conduct five variant experiments to demonstrate the rationale of our multi-stage tuning. Following LLaVA-Med (Li et al. 2024), we samely use stage I to align visual features to text embedding space. Variants (b) and (c) indicate that having alignment is more beneficial for fine-tuning in stage IV. Variants (c) and (d) demonstrate that using the E_{vl} from stage II as the initial weights for stage IV helps the model focus more on VL tasks. Similarly, variants (c) and (e) show that using the E_{ground} from stage III as the initial weights for stage IV help the model focus more on VL tasks. Lastly, the tuning strategy in stage IV as per variant (f), compared to variant (b), allows the model to better balance image-level and pixel-level tasks.

Effect of Top-k. We explored the impact of using top-1 and top-2 routing on model performance. Utilizing top-2 in our experiments implies equivalence to a dense model (where each expert processes all tokens) since we are using only two experts. The last two rows of Table 4 indicate that a dense model is less effective than a sparse activated model.

Effect of Capability Factor. We examined the impact of the Capacity Factor (CF). At CF=1, each expert handles up to half the tokens, risking information loss due to their prior knowledge. At CF=2, experts can process all tokens, leading to noise and redundancy. Empirical evidence suggests CF=1.5 is optimal, balancing the reduction of information loss and noise in token distribution.

Conclusion

In this paper, we present MedPLIB, a multimodal large language model with pixel-level insight for biomedicine. MedPLIB features flexible inputs and outputs, thereby supporting multiple tasks and creating a more versatile and patient-friendly MLLM. To achieve the mentioned targets, we have made efforts on both the model and data levels. On the model level, we introduce a three-layer architecture and a novel MoE training strategy within MLLMs that incorporates expert prior knowledge. On the data level, we introduce the MeCoVQA, which comprises an array of 8 modalities for answering complex medical imaging questions, understanding image regions, and pixel grounding. Experimental results indicate that MedPLIB has achieved state-of-the-art outcomes on the OmniMedVQA benchmark and MeCoVQA test sets. Moreover, MedPLIB has demonstrated encouraging performance in its zero-shot ability for pixel-level grounding.

References

Abacha, A. B.; Hasan, S. A.; Datla, V. V.; Liu, J.; Demner-Fushman, D.; and Müller, H. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. *CLEF (working notes)*, 2(6).

Ben Abacha, A.; Sarrouti, M.; Demner-Fushman, D.; Hasan, S. A.; and Müller, H. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chen, J.; Guo, L.; Sun, J.; Shao, S.; Yuan, Z.; Lin, L.; and Zhang, D. 2024. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1110–1119.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.

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.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

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. 2021. Towards Visual Question Answering on Pathology Images. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, volume 2.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

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.

Hu, Y.; Wang, Q.; Shao, W.; Xie, E.; Li, Z.; Han, J.; and Luo, P. 2023. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4067–4077.

Huang, X.; Li, H.; Cao, M.; Chen, L.; You, C.; and An, D. 2024. Cross-Modal Conditioned Reconstruction for Language-guided Medical Image Segmentation. *arXiv preprint arXiv:2404.02845*.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

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.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

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. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2023b. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; and Yuan, L. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

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, F.; Zhu, T.; Wu, X.; Yang, B.; You, C.; Wang, C.; Lu, L.; Liu, Z.; Zheng, Y.; Sun, X.; et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1): 226.

Liu, G.; Li, P.; Zhao, Z.; He, J.; He, G.; and Zhong, S. 2023b. Cross-Modal Self-Supervised Vision Language Pre-training with Multiple Objectives for Medical Visual Question Answering. *Authorea Preprints*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Long, Z.; Killick, G.; McCreadie, R.; and Camarasa, G. A. 2023. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval. *arXiv preprint arXiv:2309.01516*.

Luo, Y.; Zhang, J.; Fan, S.; Yang, K.; Wu, Y.; Qiao, M.; and Nie, Z. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. 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. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023b. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18155–18165.

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.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Zhang, A.; Fei, H.; Yao, Y.; Ji, W.; Li, L.; Liu, Z.; and Chua, T.-S. 2024. VPGTrans: Transfer visual prompt generator across LLMs. *Advances in Neural Information Processing Systems*, 36.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

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, B.; Jin, P.; Ning, M.; Lin, B.; Huang, J.; Song, Q.; Pan, M.; and Yuan, L. 2024. LLMBind: A Unified Modality-Task Integration Framework. *arXiv preprint arXiv:2402.14891*.

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*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.