MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output

Yanyuan Chen[1†], Dexuan Xu[2†], Yu Huang[3*], Songkun Zhan[1], Hanpin Wang[2]
Dongxue Chen[4], Xueping Wang[4], Meikang Qiu[5], Hang Li[6*]

[1]School of Software & Microelectronics, Peking University, [2]School of Computer Science, Peking University
[3]National Engineering Research Center for Software Engineering, Peking University
[4]Peking University Sixth Hospital, [5] Augusta University, [6] Peking University First Hospital

Abstract

*Currently, medical vision language models are widely used in medical vision question answering tasks. However, existing models are confronted with two issues: for input, the model only relies on text instructions and lacks direct understanding of visual clues in the image; for output, the model only gives text answers and lacks connection with key areas in the image. To address these issues, we propose a unified medical vision language model **MIMO**, with visual referring **M**ultimodal **I**nput and pixel grounding **M**ultimodal **O**utput. MIMO can not only combine visual clues and textual instructions to understand complex medical images and semantics, but can also ground medical terminologies in textual output within the image. To overcome the scarcity of relevant data in the medical field, we propose **MIMOSeg**, a comprehensive medical multimodal dataset including 895K samples. MIMOSeg is constructed from four different perspectives, covering basic instruction following and complex question answering with multimodal input and multimodal output. We conduct experiments on several downstream medical multimodal tasks. Extensive experimental results verify that MIMO can uniquely combine visual referring and pixel grounding capabilities, which are not available in previous models. Our project can be found in* https://github.com/pkusixspace/MIMO.

1. Introduction

Currently, medical vision-language models (MVLMs) equip large language models (LLMs) with visual perception modules for image understanding and leverage the powerful capabilities of LLMs to comprehend complex textual

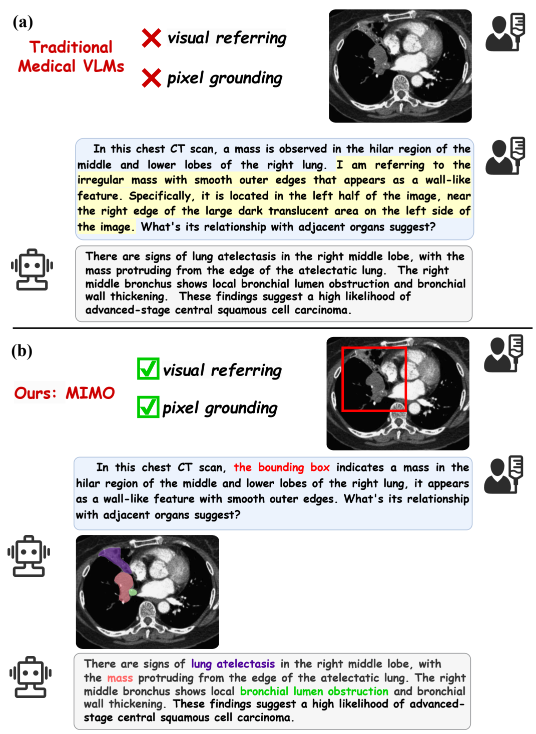


Figure 1. An illustration of visual referring and pixel grounding capabilities.

queries [7, 14, 19], demonstrating outstanding performance in medical visual question answering (MVQA). However, current models mainly use textual instructions to query the entire image, and subsequently output textual responses for image-related question answering tasks. While this approach enables a rich variety of queries and responses on the text side, it exhibits significant shortcomings in MVLM: unlike general domains, most medical terminologies typically involve specialized expressions. The limitations of text-based input and output may lead to expressions that are either overly concise or excessively complicated.

†Equally contributing first authors.
*Yu Huang and Hang Li are corresponding authors.

Figure 1 provides an illustrative example. For questions related to specific areas of medical images, accurately describing the region of interest through language is challenging (shown in Figure 1(a)). Visual Question Answering, unlike unimodal question answering, requires close integration with specific image information. Although professional physicians can use anatomical terms (e.g., "*the hilar region of the middle and lower lobes of the right lung*") to indicate these areas, however, in practical applications, the imaging findings of different patients may exhibit individual variations, further complicating the description. MVQA heavily relies on visual interpretability, even communication among specialized doctors is centered around image referring and grounding (shown in Figure 1(b)).

Inspired by the above gap, in this paper, we highlight the following two capabilities of the MVLM model: (i) visual referring in question inputs, and (ii) pixel grounding in answering outputs. Visual referring enables the model to assist users in articulating questions by combining text and visual prompts to specify a particular region [17, 30, 31], whereas pixel grounding requires the model to align semantic descriptions with segmentation masks within the image [24, 33, 37].

To construct the aforementioned MVLM system, two major challenges exist: (i) Lack of a unified model architecture. Due to design limitations, existing MVLM architectures [7, 14, 19] do not support multimodal input with visual prompts and multimodal output with segmentation masks. (ii) The scarcity of multimodal medical datasets that support these capabilities. Existing medical datasets [7, 14, 28] focus on either visual question answering or segmentation tasks, without referring and grounding annotations simultaneously.

Targeting these two challenges, we propose **MIMO**, a unified medical vision language model with visual referring **M**ultimodal **I**nput and pixel grounding **M**ultimodal **O**utput. First of all, MIMO models all visual prompts as embeddings in the same space as the image features. To effectively facilitate cross-modal alignment of these multimodal inputs, we propose a Multi-modal Input Aligner to bridge the modality gap. Then, MIMO leverages LLM as foundation, and decodes the segmentation tokens to obtain grounding masks, where these segmentation tokens are linked to semantic phrases in LLM' natural language output. Equipped with above methods, MIMO can seamless integrate natural language responses with medical entity segmentations. To train MIMO, we propose a medical multi-modal referring and grounding dataset **MIMOSeg** with 895K samples. MIMOSeg is constructed from four different perspectives, covering basic instruction following and complex question answering with multimodal input and multimodal output. We conduct experiments on several downstream medical multimodal tasks. Extensive experimental results verify that MIMO can uniquely combine visual referring and pixel grounding capabilities, which are not available in previous models. To the best of our knowledge, MIMO is the first work to simultaneously integrate visual referring and pixel grounding in medical vision language models.

Our contributions are as follows:

(1) We present **MIMO**, the first MVQA model that integrates visual referring multimodal input and pixel grounding multimodal output, specifically designed to address the unique challenges of medical image understanding and textual reasoning.

(2) We propose a comprehensive dataset **MIMOSeg** containing 895K samples, which is designed to facilitate the training of MIMO. This dataset provides a diverse range of scenarios, ensuring that the model can learn to handle both basic instructions and complex medical queries with different modalities of input.

(3) We conduct extensive held-in and held-out experiments to evaluate the performance of MIMO across various medical multimodal tasks. Experimental results verify that MIMO can uniquely combine visual referring and pixel grounding capabilities, which are not available in previous models.

2. Related Work

Multimodal Large Language Models (MLLMs). Inspired by the powerful text generation capability of LLMs, researchers are exploring ways to extend the capabilities of LLMs to the visual domain, thus advancing the development of MLLMs [2, 4, 6, 18]. Existing research on MLLMs aims to connect visual encoders with large language models. The visual encoder is employed to perceive visual information, and the LLM is used to understand semantic information and generate text. Flamingo [2] uses a visual encoder to extract visual embeddings, a resampler module to connect the visual encoder to a frozen language model, and multi-layer cross-attention module to fuse multi-modal features. BLIP2 [15] uses the Q-Former module to connect the frozen LLM and visual encoder. MiniGPT-4 [6] and LLaVA [18] freeze the parameters of the vision encoder and LLM, and only optimize a trainable linear projector to connect the vision and language layers during the instruction tuning phase.

MLLMs for referring and grounding. In recent investigations, works such as Ferret [30], GLaMM [23] and Ferret-v2 [32], closely resemble ours as they explore to enable MLLMs for image referring and grounding. Ferret designs a spatial-aware visual sampler that supports free-form shapes of the referential region. For grounding, the coordinates of the bounding box are represented in natural language numerical form in the responses, followed by the corresponding text output. GLaMM supports pixel-level grounding segmentation and expands its vocabulary with

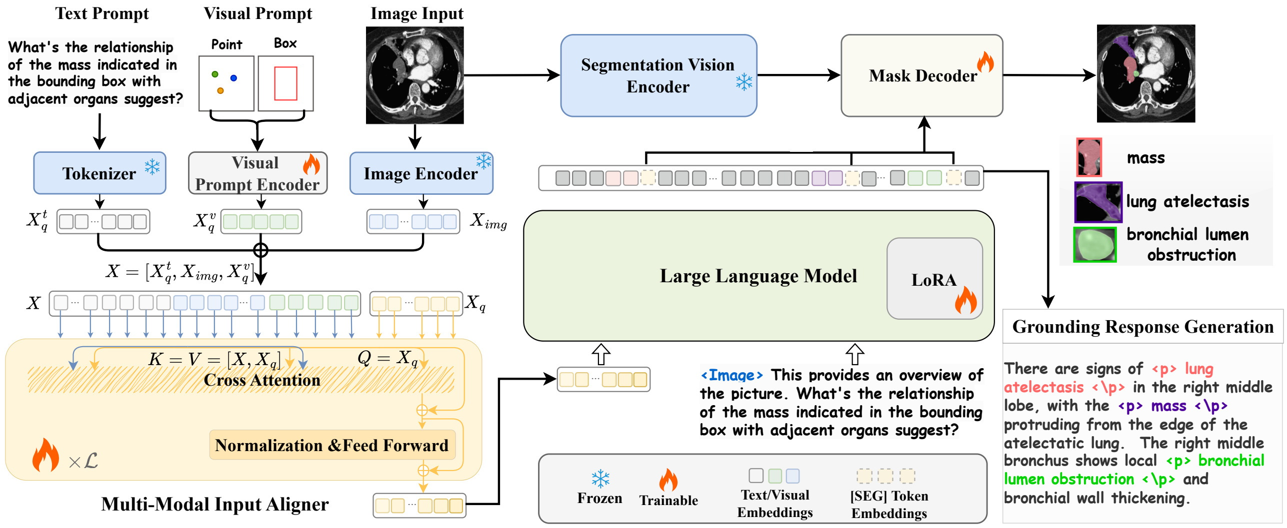


Figure 2. The overall architecture of MIMO. The model consists of an image encoder, a visual prompt encoder, a LLM, a segmentation vision encoder and a mask decoder. While accepting optional visual input, the model can also provide visual segmentation results associated with the medical entities in the text response.

special tokens to represent the input bounding box. Subsequently, the input tokens are replaced with RoI features extracted from the CLIP global image encoder layers, thereby enabling the use of specific regions as input.

MLLMs for Medical VQA. Developments in multi-modal large-scale language models have significantly impacted the biomedical field [10, 29, 35, 36]. LLaVA-Med [14] organizes biomedical image-text pairs from PubMed Central and uses GPT-4 [8] to autonomously generate instruction data for multi-modal biomedical tasks. By constructing a large set of biomedical image and text pairs, LLaVA is fine-tuned to align image-text tags to the biomedical domain. Based on the framework OpenFlamingo [3] and a pre-trained medical dataset, Med-Flamingo [19] brings Flamingo's in-context learning capability into the medical field. This adaptation allows users to tailor response formats using few-shot prompts, such as asking for an explanation of an answer. HuatuoGPT-Vision [7] leverages the large-scale dataset PubMedVision containing 1.3 million samples to train the foundational MLLM. Furthermore, MedTrinity-25M [28] employs an automated pipeline to introduce local annotations for regions of interest (ROIs), which include bounding boxes and segmentation masks. In the dataset construction strategy, different pre-trained models are utilized to annotate ROIs for the images, ensuring that each sample contains triplets of {image, ROI, description }.

3. Our Model: MIMO

3.1. Preliminary

Given an input medical image $I \in \mathbb{R}^{H\times W\times 3}$ and an input query Q, our first goal is to obtain a response R that contains a text response and seamlessly integrates segmentation masks directly tied to corresponding concepts. Let $T = [t_1, t_2, ..., t_n]$ be the text response in the form of a sequence of word tokens, then the final output R will be:

$$
\begin{aligned}
R = &\{t_j \mid j = 1, 2, \dots, n\} \\
&\cup \{< c_i, s_i >\mid s_i \in S, c_i \in T, i = 1, 2, \dots, m\}, \quad (1)
\end{aligned}
$$

where $\{t_j \mid j = 1, 2, \dots, n\}$ denotes the text response T, and $C = \{c_1, c_2, \dots, c_m\} \subseteq T$ is the set of medical concepts mentioned in the text response, $S = \{s_1, s_2, ..., s_m\}$ is the segmentation mask associated with each medical concept and $m \leq n$.

Optionally, additional visual prompts can be introduced to specify certain parts or concepts, helping to clarify the issue alongside the text. This means that the input query Q has two forms: the text query Q_t and the additional visual prompt Q_v.

3.2. Model Architecture

As shown in Figure 2, MIMO mainly consists of (i) an image encoder, (ii) a visual prompt encoder, (iii) a multi-modal input aligner, (iv) a LLM, (v) a segmentation vision encoder and (vi) a mask decoder. This model design accommodates both multimodal input and output formats, as defined in Section 3.1.

3.2.1 Visual Referring Multimodal Input

For an input medical image $I \in \mathbb{R}^{H\times W\times 3}$, a CLIP visual encoder ViT-H/14 [21] is used, combined with a projection layer to produce the image embeddings $X_{img} \in \mathbb{R}^{l_1\times d}$. For the input text query Q_t, the pre-trained tokenizer of LLM is used and projected to the text embeddings $X_q^t \in \mathbb{R}^{l_2\times d}$. In

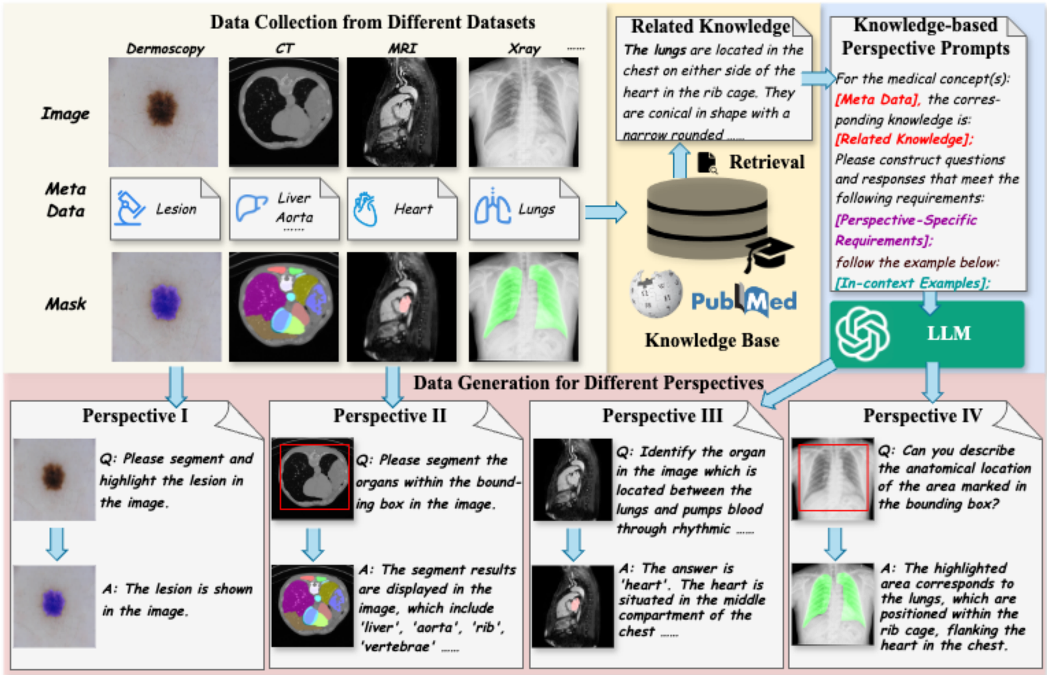


Figure 3. The construction pipeline of MIMOSeg. The pipeline inclues data collection, knowledge retrieval, prompt construction and QA generation. The bottom of the figure shows example data from four perspectives of MIMOSeg.

addition, according to practical use, we consider two types of sparse visual prompts: points and bounding boxes. The prompt encoders adapt these additional visual prompts Q_v to the same feature space as the image feature. Similar to the prompt encoder of SAM [9], points and boxes are encoded using positional encodings summed with learned embeddings for each prompt type. The corresponding embedding representation are then projected to $X_q^v \in \mathbb{R}^{l_3 \times d}$. The above image embeddings, text embeddings, and visual prompt embeddings will be concatenated into a unified multimodal feature.

Multi-modal Input Aligner. We design a multimodal input aligner to extract instruction-guided information from multi-modal features. The multi-modal input (X_{img}, X_q^t, X_q^v) interacts with the learnable query embedding X_q through cross-attention, encouraging the extraction of image features that the textual and visual prompts focus on. Therefore, LLM receives visual information that helps follow textual and visual instructions. The output of Multi-modal input aligner contains the final learned instruction-guided query embedding X_q, which is then passed through a linear projection before being fed into the LLM.

3.2.2 Language Grounding to Pixel Segmentation

We use Vicuna [34] as our language model, which is a decoder-only large language model based on LLaMA [26]. We introduce a pair of grounding tokens <p> and </p> in the vocabulary of the model to indicate the start and end of groundable medical entities, and use a special token <SEG> immediately following the groundable entities to indicate potential segmentation requests. For example, a sentence in the final output R from the model can be represented as "<p>*The central vein of the adrenal medulla*<SEG></p> *is located in the* <p>*adrenal medulla*<SEG></p> *and is a rare type of blood vessel. Its structure is different from other veins, in which the* <p> *smooth muscle* <SEG></p> *of the membrane is arranged in obvious longitudinal bundles.*"

In this way, the output embedding r_{seg} corresponding to <SEG> tokens from the last layer of the LLM can be extracted and mapped to the feature space of the decoder through a projection layer. Then, the projected hidden states will be used by the mask decoder to generate binary segmentation masks $\mathcal{M}$. The process can be formulated as:

$$\mathcal{M} = \mathcal{V}(\mathcal{G}(I), proj(r_{seg})), \tag{2}$$

where $\mathcal{V}$ is the mask decoder, $\mathcal{G}$ is the segmentation mask encoder, and the input image will pass through $\mathcal{G}$ to obtain image features as additional input of the mask decoder and $proj$ represents the projection layer. We follow SAM to employ a pretrained SAM encoder as $\mathcal{G}$ and use the pretrained SAM decoder to instantiate $\mathcal{V}$.

3.3. Training Strategy

We apply cross-entropy loss $\mathcal{L}_{text}$ for text generation. To compute the segmentation mask loss $\mathcal{L}_{mask}$, we use a combination of per-pixel binary cross-entropy (BCE) loss $\mathcal{L}_{bce}$ and DICE loss $\mathcal{L}_{dice}$. The final loss calculation is a linear combination of the language modeling loss $\mathcal{L}_{text}$ and these mask-related losses, which can be formulated as :

$$L_{total} = \lambda_1 L_{text} + \lambda_2 L_{bce} + \lambda_3 L_{dice}, \tag{3}$$

where λ_1, λ_2 and λ_3 are the corresponding hyperparameters.

4. Our Dataset: MIMOSeg

4.1. Construction Perspectives

To guide the construction of the MIMOSeg dataset, we summarize four key perspectives, each aiming to enhance the utility of the dataset in various pixel-level grounding tasks.

Perspective I: Language-Guided Segmentation. This perspective represents the basic instruction-following ability of the model. The model receives an image and a language instruction, producing segmentation masks, with no visual prompt provided. From this perspective, the model needs to build the ability to recognize medical entities and achieve the mapping from text to segmentation. An example instruction is: "*Please segment out the heart in the picture.*"

Perspective II: Visual Prompt Perceiving. This perspective represents the basic visual prompt perception ability of the model. The model receives an image, a language instruction, and a visual prompt (box or point), producing segmentation masks. From this perspective, the model needs to perceive and understand the given visual prompts, comprehend the semantics of visual entities, and establish a mapping from segmentation to text. An example instruction is: "*Please segment all the organs in the box in the picture.*"

Perspective III: Responses with Segmentation Aligning. This perspective deals with complex tasks involving question-answering. The model is required to reason about the image and generate a mask that aligns with both the question and visual segmentation, blending language understanding with pixel-level image analysis. From this perspective, the capabilities of the model are more challenging. An example question is: "*Could you identify the organ in the image that has rhythmic sequences of contractions that propagate through the atrioventricular node and conduction system?*"

Perspective IV: Visual Prompt Assisted Questioning. This perspective deals with answering a question with the assistance of a visual prompt. The visual prompt aids in directing attention, while the model needs to generate both an answer and a corresponding segmentation mask based on the question. From this perspective, the model needs to incorporate visual prompts to understand the question based on complex reasoning. An example question is: "*Can you provide details about the physiological functions of the area within the bounding box that appears to be conical and situated in the chest?*"

The four perspectives in the construction of the MIMOSeg Dataset provide a comprehensive framework for addressing different aspects of pixel-level grounding tasks. The ground-segmentation perspectives (I & II) allow models to focus on accurately delineating anatomical structures based on either text or visual cues, while the QA-driven perspectives (III & IV) simulate more complex, interactive clinical scenarios, where both segmentation and reasoning are required. This multi-perspective approach enriches the applicability of the dataset and guides the subsequent construction pipeline.

4.2. Construction Pipeline

The complete construction pipeline is illustrated in Figure 3, which includes data collection, knowledge retrieval, prompt construction and Q&A generation.

Data Source. In order to solve the problems under the above perspectives, our proposed MIMOSeg Dataset is constructed based on a wide range of publicly available segmentation datasets, with about 1 million medical pixel-level grounding samples, including data from 8 different modalities such as CT, X-rays, fundus images, and pathology sections. More details can be found in Appendix B.1. We divide these datasets to meet the construction needs of different perspectives. For each perspective, we select a sufficient number of single-label and multi-label datasets, ensuring comprehensive coverage of image modalities. The specific division is detailed in Appendix B.3.

Knowledge Retrieval. Unlike general fields, the medical field is highly dependent on knowledge, and most medical entities have corresponding definitions. Therefore, we manually constructed a knowledge base that covers all segmentation labels via some easy-access knowledge sources [1] [2]. The knowledge includes common information on disease causes, treatment methods, common symptoms, as well as the locations and functions of organs. For the metadata of the input data, we can retrieve corresponding knowledge for each label from the knowledge base.

Data Generation for different perspectives. We directly construct corresponding instruction templates and response templates for Perspective I & II. For Perspective III & IV, we design knowledge-based perspective prompts to make GPT-4o [1] generate meaningful questions and answers in a tone as if it could see the image (even though it only has access to the text). In the prompt, each perspective will correspond to different perspective-specific requirements. Perspective III focuses on identifying organs or lesions through complex questions and providing analytical reasons, while Perspective IV emphasizes visual analysis and content understanding based on visual prompts. The knowledge corresponding to the label of each image

[1] Wikipedia https://en.wikipedia.org/wiki/

[2] UMLS https://www.nlm.nih.gov/research/umls/

Table 1. The results of comparison with previous studies on four perspectives in segmentations. Best and second-best results are shown in **bold** and underline, respectively. × means that the model does not support text instruction following.

Models	SAM-b [12]	SAM-h [12]	SAM-Med [9]	GLaMM [24]	MIMO(w/o Aligner)	MIMO
Perspective I						
mIoU ↑	×	×	×	0.556	0.607	**0.639**
AP50 ↑	×	×	×	**0.324**	0.301	0.316
F1 ↑	×	×	×	**0.507**	0.448	0.457
Perspective II						
mIoU ↑	0.568	0.571	0.494	0.496	0.622	**0.665**
AP50 ↑	0.125	0.132	0.097	0.137	0.185	**0.279**
F1 ↑	×	×	×	0.279	0.314	**0.406**
Perspective III						
mIoU ↑	×	×	×	0.421	0.468	**0.531**
AP50 ↑	×	×	×	0.133	0.205	**0.302**
F1 ↑	×	×	×	0.253	0.340	**0.427**
Perspective IV						
mIoU ↑	0.573	0.583	0.503	0.564	0.526	**0.586**
AP50 ↑	0.363	**0.404**	0.300	0.267	0.253	0.309
F1 ↑	×	×	×	0.422	0.401	**0.470**

in the aforementioned knowledge base will be input into the prompt. For multi-label tasks, the generated questions and answers must consider each label and their interrelationships. We also manually construct relevant in-context examples to enhance the generation of Q&A pairs. The specific prompts are detailed in Appendix E.2. Examples of Q&A content for the four types of perspectives are shown in Appendix C.

5. Experiments and Analysis

5.1. Experimental Setup

Datasets. Since MIMOSeg is the first dataset with both visual referring and pixel grounding, we conduct held-in experiments on the MIMOSeg test set. We show experimental results under four perspectives to compare the capabilities of the models more reasonably. Meanwhile, to evaluate the generalization ability of our model, we select 6 held-out datasets for zero-shot experiments, including 3 medical VQA datasets Slake [16], RadVQA [13], PathVQA [11] and 3 medical segmentation datasets X-ray [9], Fundus [9], Skin Lesion [22].

Metrics. In the held-in experiments, we report BLEU-4 [20], ROUGE-L [25], METEOR [5] for text generation evaluation, AP50 and mIOU for mask segmentation, and F1 Score for mask-to-entity correspondence accuracy (refer to Appendix D for details). In the held-out experiments, for the VQA task, we evaluate the accuracy of the model's responses. The experimental settings remain the same as HuatuoGPT-Vision. For the segmentation task, we use the same settings as the held-in experiments.

Implementation Details. We use Vicuna LLM with 7B parameters as the default large language model, and the multi-modal input aligner is randomly initialized. After initialization, MIMO is trained on the aforementioned MIMOSeg and LLaVA-Med VQA [14] dataset. Specifically, the four perspectives of MIMOSeg and the LLaVA-Med VQA dataset are mixed for training with a ratio of $1:2:2:1:1$ for 3 epochs. More implementation details can be found in Appendix D.

5.2. Comparison with Previous Studies

We conduct comparative experiments of our method with existing SAMs, existing medical VLMs, and general VLMs. For existing SAMs and medical VLMs, we directly conduct zero-shot experiments. For general VLMs, we select a base model that supports multimodal input and output, and fine-tune it using MIMOSeg. It should be noted that since there is no model in the general field combines both visual reference multimodal input and pixel-grounded multimodal output, we only select the most similar model and make appropriate modifications to better suit our task. The experimental results are shown in Table 1 to Table 4.

Held-in experiments. We first evaluate MIMO on the held-in dataset, summarizing results from four different perspectives in Table 1 and Table 2. For the segmentation performance in tasks involving visual prompts (Perspective II & IV), we compare MIMO's performance with SAM and SAM-Med. Since SAM only supports segmentation mask outputs and does not provide textual responses, its F1 scores are marked as ×. As shown in Table 1, MIMO demonstrates significantly better performance compared to baseline methods.

In complex scenarios (Perspective III & IV), MIMO outperforms our reproduced GLaMM model, indicating the architectural advantages of MIMO. We also present experi-

Table 2. The results of comparison with previous studies on four perspectives in question answering. Best and second-best results are shown in **bold** and underline, respectively.

Models	**GPT-4o** [1]	**LLaVA-Med** [14]	**HuatuoGPT-Vision** [7]	**GLaMM** [24]	**MIMO**(w/o Aligner)	**MIMO**
Perspective II						
ROUGE-L ↑	0.169	0.120	0.070	**0.436**	0.317	0.335
BLEU-4 ↑	0.002	0.005	0.001	**0.293**	0.171	0.181
METEOR ↑	0.053	0.077	0.062	**0.304**	0.208	0.253
Perspective III						
ROUGE-L ↑	0.146	0.269	0.278	0.339	**0.650**	0.581
BLEU-4 ↑	0.007	0.129	0.064	0.150	**0.564**	0.475
METEOR ↑	0.073	0.163	0.211	0.198	**0.392**	0.334
Perspective IV						
ROUGE-L ↑	0.223	0.215	0.222	0.380	0.387	**0.397**
BLEU-4 ↑	0.061	0.046	0.034	0.240	**0.263**	**0.263**
METEOR ↑	0.159	0.127	0.156	0.216	0.218	**0.222**

Table 3. Comparison with previous work on the held-out segmentation datasets. Best and second-best results are shown in **bold** and underline, respectively. × means that the model does not support text instruction following.

Models	**X-ray**				**Fundus**				**Skin Lesion**			
	w/o bbox		with bbox		w/o bbox		with bbox		w/o bbox		with bbox	
	AP50	mIoU	AP50	mIoU	AP50	mIoU	AP50	mIoU	AP50	mIoU	AP50	mIoU
SAM-b [12]	×	×	0.971	0.855	×	×	0.923	0.766	×	×	0.952	0.851
SAM-h [12]	×	×	**0.989**	0.835	×	×	0.973	0.865	×	×	0.982	0.861
SAM-Med [9]	×	×	0.955	0.877	×	×	0.678	0.662	×	×	0.931	0.848
GLaMM [24]	0.335	0.554	0.859	0.882	**0.940**	**0.870**	0.946	0.874	0.458	0.526	0.719	0.723
MIMO (w/o Aligner)	0.313	0.490	0.822	0.813	0.915	0.852	0.951	**0.883**	0.747	0.638	**0.985**	0.861
MIMO	**0.507**	**0.718**	**0.989**	**0.883**	**0.940**	0.863	**0.988**	0.881	**0.787**	**0.647**	**0.985**	**0.871**

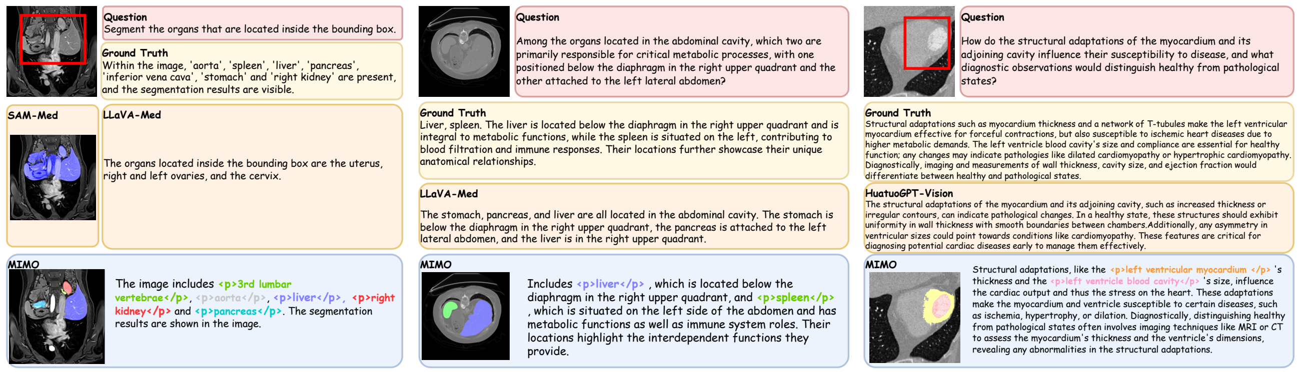
mental results using medical VLMs and GPT-4o to evaluate the text responses. Given that MIMOSeg's Q&A design focuses on specific entities within images, the four tasks based on distinct perspectives are challenging for traditional VLMs (refer to Figure 4 and Appendix F for qualitative analysis). The enhanced evaluation results of MIMO further highlight the significance of integrating unified referring and grounding capabilities in MVQA models.

Held-out experiments. Table 3 and Table 4 show the held-out experimental results. In the segmentation task, we perform segmentation with and without bounding boxes. Since SAM and SAM-Med only support visual prompt input, their results in the table w/o bbox is ×. From Table 3, we can see that MIMO achieves the best results in both types of segmentation. The experimental results show that MIMO has good generalization ability in the case of text instruction segmentation; in the case of segmentation with visual prompts, MIMO's segmentation effect also outperforms the previous SAM and SAM-Med models. In the VQA task, we conduct experiments on 3 public benchmarks. We choose closed-ended questions in the benchmarks to more accurately evaluate the accuracy of the generated results. Among them, HuatuoGPT-Vision uses more datasets for training. Compared with the remaining models, our model achieves the best accuracy on Slake and VQA-RAD. At the same time, MIMO also has the highest average accuracy on the three datasets.

5.3. Ablation Studies

Ablation of Multi-modal Input Aligner. Comparing the results of MIMO and MIMO (w/o aligner) in Table 1 to Table 3, MIMO shows higher results in most cases, especially in the case of inputs with bounding boxes such as Perspective II & IV. These results prove the effectiveness of our designed aligner. The multimodal aligner can get more accurate content understanding when the input has visual prompts.

Ablation of Training Data Ratio. The amount of data is an important factor affecting the experiment. To examine the impact of data quantity on model performance, we compare the evaluation results under different data ratios. Experiments are conducted using 25%, 50%, 75% and 100% of the training data. The experimental results are shown in Figure 5. Observing Figures 5(a), (b) and (c), we can find that as the amount of data increases, the model's mIoU, AP50, F1 and other indicators are steadily increasing. In Figure 5(d), for VQA-RAD and PathVQA, the result also shows a similar trend. These experimental results prove



(a) Perspective II (b) Perspective III (c) Perspective IV

Figure 4. Qualitative analysis of experimental results from three perspectives. We compare the results with LLaVA-Med and HuatuoGPT-Vision. MIMO can generate segmentation masks with relevant medical entities while outputting the answer.

the positive correlation between the model effect and the amount of data, and also prove that the MIMOSeg dataset can bring positive effects to multimodal medical tasks.

Table 4. Comparison with previous work on the held-out medical visual question answering datasets. The gray data represents that the model uses a larger amount of data for training and is for reference only.

Models	VQA-RAD	SLAKE	PathVQA	Avg.
HuatuoGPT-Vision [7]	**63.8**	**74.5**	**59.9**	66.1
Med-Flamingo [19]	45.4	43.5	54.7	47.9
RAD-FM [27]	50.6	34.6	38.7	41.3
LLaVa-Med [14]	51.4	48.6	**56.8**	52.3
Qwen-VL-Chat [4]	47.0	56.0	55.1	52.7
MIMO	**58.8**	**57.0**	52.4	**56.1**

5.4. Qualitative Analysis

To compare the capabilities of MIMO with existing medical vision language models, we conduct case studies on questions from different perspectives. Figure 4(a) shows the recognition and segmentation results of the model for organs in the bounding box under perspective II. It can be seen that SAM-Med can perform segmentation, but it cannot achieve the correspondence between the segmentation results and the labels. LLaVA-Med can answer the question, but due to the lack of visual association, the organ recognition results are wrong. MIMO can give segmentation results while answering the question, and has higher recognition accuracy. Figure 4(b) shows the reasoning of the model on organ function under perspective III. LLaVA-Med only answers partial organs, while MIMO answers all the correct organs and gives the corresponding segmentation results. Figure 4(c) shows the knowledge question answering results with the visual prompt under perspective IV. HuatuoGPT-Vision and MIMO both answer based on the internal knowledge of the model. However, MIMO can more intuitively give the corresponding segmentation results of the key medical entities in the answer.

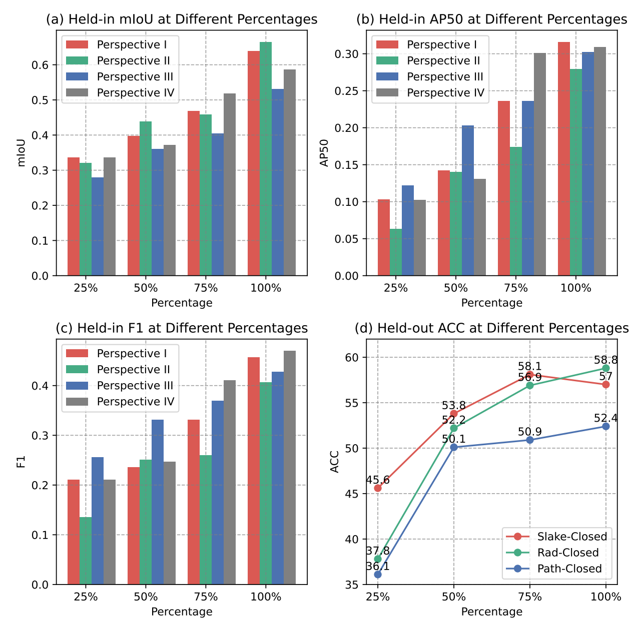


Figure 5. Held-in and held-out experimental results with different ratios of training data.

6. Conclusion

In this paper, we present MIMO, the first medical multimodal large language model capable of simultaneously accepting visual prompts and generating text responses intertwined with grounded segmentation masks. To advance research and model development, we constructed MIMOSeg dataset, a large-scale, medical multimodal referring and grounding dataset, consisting of 895K samples.

Limitations. For the Q&A generation approach used in constructing MIMOSeg, it is capable of generating the questions and answers only on the basis of the knowledge base. Consequently, its scope is bound by the given knowledge, limiting comprehensive coverage of information available for data generation.

Acknowledgment

This paper was supported by National Key R&D Program of China (No. 2023YFC3502902, 2021YFF1201100). We thank Huamin Zhang's team from Institute of Basic Theory of Chinese Medicine, China Academy of Chinese Medical Sciences for assistance.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 8

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[7] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 1, 2, 3, 7, 8

[8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 4, 6, 7

[10] Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. 2023. 3

[11] Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 6

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6, 7

[13] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 6

[14] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1, 2, 3, 6, 7, 8

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[16] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 6

[17] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[19] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1, 2, 3, 8

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[22] Venkatesan Rajinikanth, Suresh Chandra Satapathy, Nilanjan Dey, Steven Lawrence Fernandes, and K Suresh Manic. Skin melanoma assessment using kapur's entropy and level set—a study with bat algorithm. In *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 1*, pages 193–202. Springer, 2019. 6

[23] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2

[24] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2, 6, 7

[25] Lin CY Rouge. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 6

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[27] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 8

[28] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. 2, 3

[29] Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, et al. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997, 2024. 3

[30] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2

[31] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 2

[32] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 2

[33] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238, 2024. 2

[34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 4

[35] Juexiao Zhou, Xiuying Chen, and Xin Gao. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *arXiv preprint arXiv:2306.10765*, 2023. 3

[36] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. 2023. 3

[37] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6587, 2019. 2