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

Supplementary Material

A. Functional Comparison with Other Models

As is shown in Table 5. Current medical multimodal large language models [2, 13, 39] face challenges in comprehensively supporting both referring and grounding capabilities. Meanwhile, some models in general domains are designed to address individual functions, including leveraging visual prompts to enhance referring capabilities [10, 14, 41, 79, 90] or generating mask segmentation outputs [37, 81, 84]. A line of works attempt to combine visual prompt inputs with mask segmentation outputs but lack grounding capabilities [92, 96]. Models [77, 89] that incorporate both referring and grounding often rely on bounding box-based localization, which may be insufficient for high-precision requirements in medical contexts. A few models [66, 95] have shown significant advancements in integrating referring and pixel-wise grounding. However, these approaches often rely on complex visual prompt encoding mechanisms and extensive general-domain datasets, making direct adaptation to the medical field challenging. Among these, our proposed MIMO adopts a concise design that integrates visual referring and pixel-level grounding capabilities. By supporting multimodal inputs and multimodal outputs, it offers an efficient and precise solution tailored to the unique demands of medical applications.

B. Data Statistics

B.1. Data Modalities

Figure 6 shows the modality statistics of MIMOSeg. MIMOSeg contains eight types of images of different medical modalities. The order from most to least is CT, MRI, Dermoscopy, PET, Endoscopy, X-Ray, Ultrasound and Fundus.

B.2. Data Source

Table 10 shows the data source of MIMOSeg. For each data source, we count the number of images, the number of masks, the modality type of the dataset and the labels of the dataset, and show which perspective the dataset is used to construct data.

B.3. Statistics of Different Perspectives

Table 6 shows the statistical results of MIMOSeg, including the number of training, validation, and test sets for each perspective. Since the amount of MIMOSeg data is very large, we divide the training, validation, and test sets in a ratio of 99:0.5:0.5.

B.4. Statistics of Zero-shot Datasets

To demonstrate the generalization ability of the model, we conduct zero-shot tests on 6 held-out datasets. Table 7 shows the statistical results of the number of each zero-shot dataset.

C. Data Analysis

Figure 7 presents several examples of MIMOSeg. As mentioned, different perspectives focus on different aspects. Perspective I focuses on the model's ability to directly follow instructions. Perspective II focuses on the model's ability to understand visual clues. Perspective III addresses segmentation associated with complex reasoning. Perspective IV focuses on complex question answering with visual clues. In the divided test and validation sets, we further manually filtered the test and validation sets for Perspective III and Perspective IV to obtain high-quality evaluation data. Specially, we manually excluded data containing questions or answers unrelated to the image content.

D. More Implementation Details

D.1. Evaluation metric

F1 Score. Inspired by [89, 101], we propose the F1 Score to quantify the grounding capability of masks aligned with medical entities. For the multimodel output $R = \{t_j \mid j = 1, 2, \ldots, n\} \cup \{< c_i, s_i > \mid s_i \in S, c_i \in T, i = 1, 2, \ldots, m\}$, we define A as the total number of entity words (c) in the generated sentence R, and B as the total number of entity words in the ground truth sentence. E represents the total number of correct prediction pairs ($\{< c, s >\}$). In this paper, a correct prediction refers to generating entity words that match those in the ground truth and producing a correct mask segmentation (i.e., IoU with the ground truth segmentation > 0.5). In F1, the precision and recall can be defined as

$$\text{Precision} = \frac{E}{A}, \quad \text{Recall} = \frac{E}{B}. \tag{4}$$

The F1 score is calculated as follows:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

D.2. Model Architecture and Training

We use Vicuna LLM with 7B parameters as the default large language model and instantiate the image encoder with ViT-H/14 CLIP. The visual prompt encoder adopts a positional encoding scheme similar to SAM. The multi-modal

Method	Multimodal Input			Multimodal Output				Medical LLM?
	Image	Text	Visual Prompt	Text	Mask	Mask-to-Phrase Grounding	Multi-Region	
LLaVA [46]	✔	✔	✗	✔	✗	✗	✗	✗
miniGPT4 [12]	✔	✔	✗	✔	✗	✗	✗	✗
mPLUG-OWL [88]	✔	✔	✗	✔	✗	✗	✗	✗
LLaMA-Adapter v2 [25]	✔	✔	✗	✔	✗	✗	✗	✗
InstructBLIP [20]	✔	✔	✗	✔	✗	✗	✗	✗
LLaVA-med [39]	✔	✔	✗	✔	✗	✗	✗	✔
HuatuoVision [13]	✔	✔	✗	✔	✗	✗	✗	✔
Med-Flamingo [2]	✔	✔	✗	✔	✗	✗	✗	✔
Caption Anything [79]	✔	✔	✔	✔	✗	✗	✗	✗
Osprey [90]	✔	✔	✔	✔	✗	✗	✗	✗
ViP-LLaVA [10]	✔	✔	✔	✔	✗	✗	✗	✗
SPHINX-V [41]	✔	✔	✔	✔	✗	✗	✗	✗
Shikra [14]	✔	✔	✔	✔	✗	✗	✔	✗
ASMv2 [80]	✔	✔	✗	✔	✗	✔/✗	✔	✗
GSVA [85]	✔	✔	✗	✔	✔	✗	✔	✗
LaSagnA [81]	✔	✔	✗	✔	✔	✔	✔	✗
LISA [37]	✔	✔	✗	✔	✔	✗	✗	✗
PSALM [96]	✔	✔	✔	✔	✔	✗	✗	✗
NextChat [92]	✔	✔	✔	✔	✔	✗	✔	✗
Ferret [89]	✔	✔	✔	✔	✗	✔/✗	✔	✗
ChatterBox [77]	✔	✔	✔	✔	✗	✔/✗	✗	✗
GroundHog [95]	✔	✔	✔*	✔	✔	✔	✔	✗
GLaMM [66]	✔	✔	✔*	✔	✔	✔	✔	✗
MIMO(ours)	✔	✔	✔	✔	✔	✔	✔	✔

Table 5. Comparison of recent multimodal large language models. ✔ indicates support, while ✗ indicates no support. ✔/✗ ASMv2 does not support pixel-wise segmentation masks with phrase grounding, using bounding boxes instead, making it unsuitable for medical applications requiring fine-grained precision. ✔/✗ Ferret and ChatterBox output bounding boxes for localization instead of masks and does not support pixel-wise grounding. ✔* In GLaMM, user-input spatial prompts are limited to bounding boxes, with a carefully designed region encoder extracting a region-of-interest representation corresponding to the box. ✔* GroundHog converts user-input spatial prompts into binary masks via SAM and uses a masked feature extractor to extract local features.

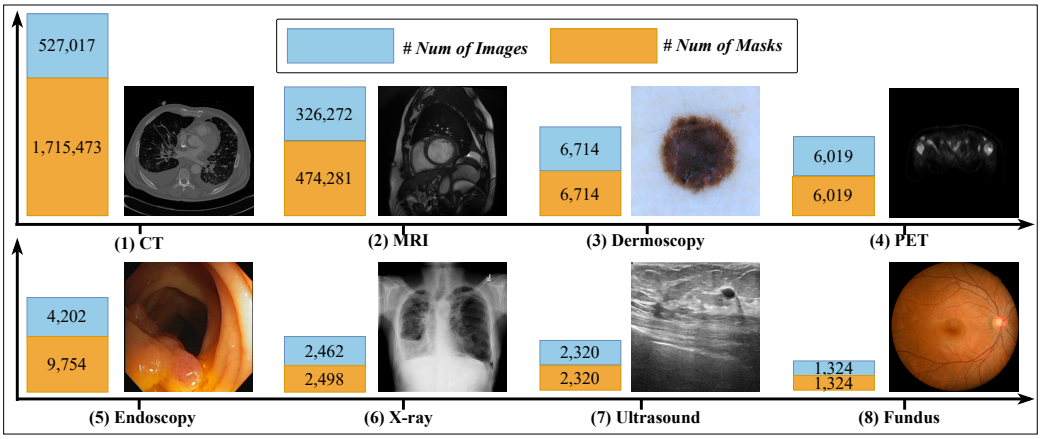

Figure 6. Modal statistics results of MIMOSeg.

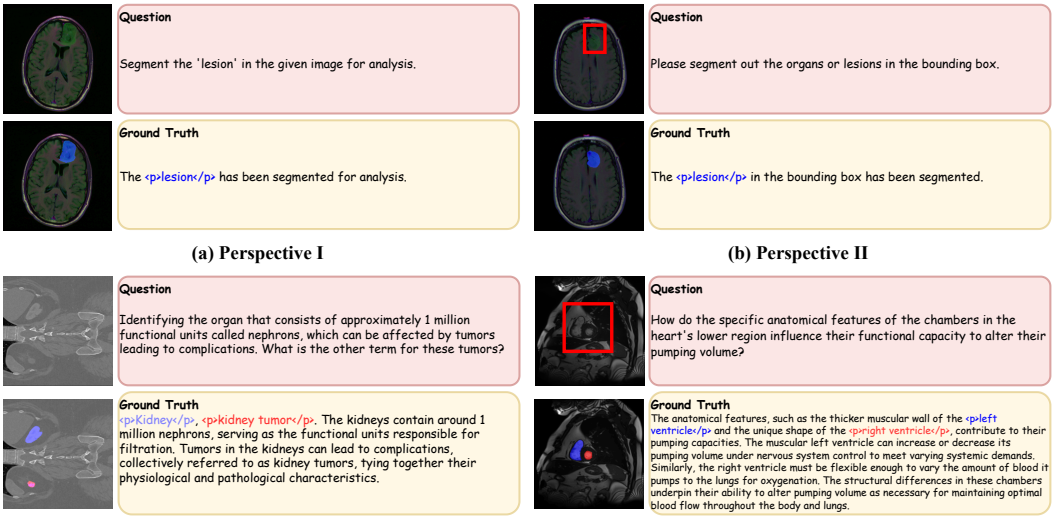

Figure 7. Several examples of MIMOSeg.

Table 6. Data Statistics for MIMOSeg, including different perspectives.

Dataset	Train	Val	Test	Total
Perspective I	249,665	2,564	2,545	254,774
Perspective II	249,698	2,552	2,334	254,584
Perspective III	181,046	726	726	182,498
Perspective IV	178,594	1,070	1,070	180,734
Total	859,003	6,912	6,675	872,590

Table 7. Data Statistics for held-out experiments.

Segmentation	X-ray	Fundus	Skin Lesion
Num	281	270	206
VQA	**VQA-RAD**	**SLAKE**	**PathVQA**
Num	272	416	3,391

input aligner is randomly initialized, while the segmentation mask encoder and mask decoder are initialized using SAM's encoder and decoder, respectively. During training, the image encoder and segmentation mask encoder remain frozen, while the visual prompt encoder, multi-modal input aligner, and mask decoder are trained. Additionally, we apply Low-Rank Adaptation (LoRA) [32] with $\alpha = 8$ to fine-tune the LLM. Our codes and pretrained models will be publicly released.

After initialization, MIMO was trained using the aforementioned MIMOSeg dataset and the 60K LLaVA-Med VQA dataset on 4 A800 GPUs for 3 epochs, which took approximately 10$\sim$12 days. The training was optimized using the Adam optimizer with a learning rate of 3e-4 and a batch size of 40. To enhance the model's capability to follow VQA instructions, the trained model is further fine-tuned for one additional epoch on the LLaVA-Med VQA dataset, with the mask decoder frozen.

E. Instructions and Prompts

E.1. Instruction Templates for Perspective I & II

E.1.1 Instruction Templates for Perspective I

To construct the instruction-following dataset in Perspective I, we design template instructions and responses. Specifically, for a given image, its corresponding masks, and the labels associated with the masks, we design segmentation instruction templates that specify the names of the medical entities to be segmented. Depending on the number of labels in each image, the responses use either single-label or multi-label response templates, as shown in Table 8.

E.1.2 Instruction Templates for Perspective II

In constructing the instruction-following dataset for Perspective II, we design question templates to trigger segmentation. Depending on the different forms of visual prompts, we formulate two types of question templates (i.e. box and point) along with corresponding answer templates, as shown in Table 9.

E.2. Prompts for Q&A generation in Perspective III & IV

E.2.1 Prompts for Q&A generation in Perspective III

Figure 8 and Figure 9 illustrate the knowledge-based prompts used to construct the Perspective III data for MI-

Instruction Template of of Perspective I
Please segment the {} in the image.
Can you identify and segment the distinct {} elements within the image?
I need the {} in the image to be categorized into individual segments.
Could you analyze the image and segment the {} into separate segments?
Can you perform an image segmentation to extract the {}?
Segment the {} in the given image for analysis.
Segment and highlight the {} in the image.
Cut out the {} from the image and display it.
Segment the {} regions in the medical image.
Response Template of Perspective I for single-label image
The image includes {}. The segmentation result is shown in the image.
The segmentation result is displayed in the image, which includes {}.
You can see the segmentation result in the image, along with {}.
The image shows the segmentation result, which includes {}.
The segmentation result in the image includes {}.
Within the image, {} is present, and the segmentation result is visible.
Response Template of Perspective I for multi-label image
The image includes {}. The segmentation results are shown in the image.
The segmentation results are displayed in the image, which include {}.
You can see the segmentation results in the image, along with {}.
The image shows the segmentation results, which include {}.
The segmentation results in the image include {}.
Within the image, {} are present, and the segmentation results are visible.

Table 8. Templates used for the Perspective I.

Instruction Template of of Perspective II for box prompt
Please segment out the organs or lesions in the bounding box.
Please identify and segment the organs or lesions within the given bounding box.
Segment the organs or lesions that are located inside the bounding box.
Can you segment out the organs or lesions found in the specified bounding box?
Please perform segmentation of the organs or lesions within this bounding box.
Segment any organs or lesions present within the provided bounding box.
Conduct segmentation of organs or lesions contained in the bounding box.
Could you segment the organs or lesions that are inside the bounding box?
Response Template of Percpective II for box prompt
The result of segmentation is {} and is shown in the image.
The outcome of the segmentation is {} and is displayed in the image.
The segmentation result is {} and is shown in the image.
The organs or lesions have been segmented and the result is {}.
The segmentation output is {} and is present in the image.
Segmentation results in {}, which is shown in the image.
Instruction Template of Perspective II for point prompt
Please segment out the organs or lesions at the specified point.
Please identify and segment the organs or lesions at the given point.
Segment the organs or lesions that are located at the specified point.
Can you segment out the organs or lesions found at the specified point?
Please perform segmentation of the organs or lesions at this point.
Segment any organs or lesions present at the provided point.
Conduct segmentation of organs or lesions at the specified point.
Could you segment the organs or lesions at the specified point?
Response Template of Perspective II for point prompt
The result of segmentation is {} and is shown in the image.
The outcome of the segmentation is {} and is displayed in the image.
The segmentation result is {} and is shown in the image.
The organs or lesions have been segmented and the result is {}.
The segmentation output is {} and is present in the image.
Segmentation results in {}, which is shown in the image.

Table 9. Templates used for the Perspective II.

MOSeg. Each prompt contains medical entity names, their corresponding knowledge, and in-context learning examples. An in-context learning example is shown in Figure 10. For images with multiple mask labels, we concatenate the knowledge associated with the entity labels and input it into the prompt, emphasizing the relationships between multiple entities.

E.2.2 Prompts for Q&A generation in Perspective IV

Figure 11 and Figure 12 present the knowledge-based prompts used to construct the Perspective IV data for MIMOSeg. Each prompt includes medical entity names and their corresponding knowledge. For images with multiple mask labels, we concatenate the knowledge associated with each entity label and input it into the prompt, focusing on the relationsHips between the entities.

F. More Experimental Results

Figures 13 and Figure 14 show the case analysis results of LLaVA-Med and HuatuoGPT-Vision under four perspectives. MIMO can provide the segmentation results of the medical entities associated with the answer while giving the correct text reply.

G. Ethics Statement

MIMO is built upon a large language model (LLM), inheriting original limitations of LLMs, such as language hallucinations, where it may generate harmful or counterfactual responses. Moreover, machines are not infallible, and there is a potential risk that the model may misinterpret user inputs or make inaccurate predictions. In high-risk medical settings, such errors could be harmful or even dangerous. Researchers and developers must be aware of the potential risks associated with the use and misuse of medical LLMs in healthcare environments and implement both automated safeguards (e.g., setting strict thresholds for diagnostic suggestions) and human interventions (e.g., training staff to recognize potential system failures). We explicitly state that the MIMOSeg dataset we release is intended solely for research purposes. Furthermore, our data collection methods comply with the terms of use and adhere to the intellectual property and privacy rights of the original authors.

Prompting GPT-4o to generate Q&A of Perspective Ⅲ for single-label image

For the medical concept: [Meta Data],
the corres-ponding knowledge is: [Related Knowledge];
Please construct questions and responses that meet the following requirements:
(1) Construct questions based solely on the given knowledge.
(2) For the questions posed, the answer must be the given medical concept, meaning the answer can only be 'the answer is [Meta Data]', so the questions need to be as detailed and precise as possible and should highly distinguish this medical concept.
(3) Questions should meet medical professionalism and be diversified around physiological functions, anatomical locations, biochemical properties, pathological features, diagnostic and therapeutic characteristics, etc.
(4) The answer to the questions must be unique, with the only correct answer being 'the answer is [Meta Data]', and any other answer being incorrect.
(5) Integrate all given knowledge to provide a systematic and comprehensive detailed explanation for the answer.\n Provide five questions and corresponding answers and explanations as your <Responce>. You can only return your <Response> as a list, follow the example below:[In-context Examples];

Figure 8. Knowledge-based Prompt for Q&A generation on single-label data in Perspective III.

Prompting GPT-4o to generate Q&A of Perspective Ⅲ for multi-label image

For the medical concepts: [Meta Data],
the corres-ponding knowledge is: [Related Knowledge];
Please construct questions and responses that meet the following requirements:
(1) Only use the given knowledge to construct questions.
(2) If possible, explore the potential relationships between [Meta Data] from medical professional perspectives such as physiological functions, anatomical positions, biochemical characteristics, pathological features, diagnostic and therapeutic characteristics, etc., and ask different questions.
(3) Questions should meet medical professionalism and be diversified around physiological functions, anatomical locations, biochemical properties, pathological features, diagnostic and therapeutic characteristics, etc.
(4) The answer to the questions must be unique, with the only correct answer being 'the answer are [Meta Data]', and any other answer being incorrect.
(5) Integrate all given knowledge to provide a systematic and comprehensive detailed explanation for the answer.\n Provide five questions and corresponding answers and explanations as your <Responce>. You can only return your <Response> as a list, follow the example below:[In-context Examples];

Figure 9. Knowledge-based Prompt for Q&A generation on multi-label data in Perspective III.

In-context Learning Example

<Knowledge>: "The duodenum is part of the digestive system, located between the stomach and the jejunum. The duodenum forms a 'C' shape around the head of the pancreas, dividing it into the superior, descending, horizontal, and ascending parts according to its course. A longitudinal fold on the posterior-medial wall of the middle part of the descending duodenum features a protrusion called the major duodenal papilla, which is the common opening of the bile duct and pancreatic duct (Wirsung duct). The circular muscle band surrounding the major duodenal papilla is called the sphincter of Oddi, which regulates the secretion of pancreatic juice and bile into the duodenum. The gallbladder's bile passes through the cystic duct, common bile duct, ampulla of Vater, and major duodenal papilla into the duodenum. The pancreas is the second largest digestive gland in the human body, with both exocrine and endocrine functions, located in the epigastrium and left hypochondrium, horizontally positioned in front of the first and second lumbar vertebrae and closely attached to the posterior abdominal wall. The pancreas is divided into three parts: head, body, and tail, with the head surrounded by the 'C' groove of the duodenum. The exocrine fluid of the pancreas, called pancreatic juice, contains large amounts of enzymes that break down and digest proteins, carbohydrates, and fats, and is essential for digestion. The blood supply of the pancreas and duodenum both come from the pancreaticoduodenal artery arc, formed by the anterior and posterior branches of the superior and inferior pancreaticoduodenal arteries. The main duct of the pancreas, known as the Wirsung duct, opens into the duodenum either through an independent opening or through a common passage, the ampulla of Vater."

<Response>: [{"question": "Which two structures of the digestive system are supplied by the pancreaticoduodenal artery arc, with one forming a 'C' shape around the other's head?","answer": "Duodenum, Pancreas.","explanation": "The duodenum is part of the digestive system, and the pancreas is the second largest digestive gland in the human body with both exocrine and endocrine functions. The duodenum forms a C-shape around the head of the pancreas, which explains their close anatomical relationship. The blood supply to both the head of the pancreas and the duodenum comes from the pancreaticoduodenal artery arc."},{"question": "Which two organs have a close anatomical relationship in the digestive system, with the exocrine fluid of one regulated by the sphincter of Oddi for secretion into the other?","answer": "Duodenum, Pancreas.","explanation": "The duodenum and pancreas have a close anatomical relationship in the digestive system. These two organs interact through the sphincter of Oddi, which regulates the secretion of the pancreatic exocrine fluid (pancreatic juice) into the duodenum. Additionally, bile secretion is also regulated by the sphincter of Oddi and is discharged through the major duodenal papilla into the duodenum, aiding in the digestion of food.

Figure 10. An in-context learning example.

Prompting GPT-4o to generate Q&A of Perspective Ⅳ for single-label image

Consider this scenario: 'There is a medical image, and the user can use a bounding box (or a point) to specify certain medical entity within the image. Then, the user poses textual questions about the entity, and the Assistant responds by providing fine-grained segmentation of the entity corresponding to the selected area along with the name of the entity, accompanied by a textual answer.' In this scenario, users can effectively use the [bounding box](or [point]), and [textual question] to specify area and clearly articulate their questions for visual interactive querying with the Assistant.\n Now, I need you to design the [textual question] and [textual answer] parts for the above scenario that meet the following requirements:
(1) Act as the questioning user to design the [textual question], and act as the Assistant to design the [textual answer].
(2) I cannot provide you with the actual image and bounding box(or point), so in the [textual question] part, you need to act as if you are viewing the image and bounding box(or point), note that the user can only enter either bbox or point, please consider different scenarios for your question
(3) The user does not know the name of the entity within the bounding box(or point), so they use the bounding box(or point) to assist in making the textual question clear; without the bounding box(or point), the question would be unclear.
(4) Although I cannot provide the actual image and bounding box(or point), I can tell you the name of the entity specified by the bounding box(or point) and provide relevant knowledge about the entity to aid in designing the textual question and answer. Absolutely do not disclose the name of the entity in the textual question.
(5) Use only the given knowledge to construct questions; do not make up information, as you cannot actually see the image and confirm its medical modality, therefore avoid using knowledge that might not match the information in the image.
(6) Focus on physiological functions, anatomical positions, biochemical characteristics, pathological features, and diagnostic treatment characteristics from a medical professional perspective.
(7) In the [textual answer], you must explain the name of the entity included in the bounding box(or point), and provide a systematic and comprehensive answer integrating all the given knowledge.
<Specified medical entities in the bounding box(or point)>: [Meta Data],
<Knowledge>: [Related Knowledge];

Figure 11. Knowledge-based Prompt for Q&A generation on single-label data in Perspective IV.

Prompting GPT-4o to generate Q&A of Perspective IV for multi-label image

Consider this scenario: 'There is a medical image, and the user can use a bounding box (or a point) to specify certain medical entity within the image. Then, the user poses textual questions about the entity, and the Assistant responds by providing fine-grained segmentation of the entity corresponding to the selected area along with the name of the entity, accompanied by a textual answer.' In this scenario, users can effectively use the [bounding box](or [point]), and [textual question] to specify area and clearly articulate their questions for visual interactive querying with the Assistant.\n Now, I need you to design the [textual question] and [textual answer] parts for the above scenario that meet the following requirements:
(1) Act as the questioning user to design the [textual question], and act as the Assistant to design the [textual answer].
(2) I cannot provide you with the actual image and bounding box, so in the [textual question] part, you need to act as if you are viewing the image and bounding box.
(3) The user does not know the names of the entities within the bounding box, so they use the bounding box to assist in making the textual question clear; without the bounding box, the question would be unclear.
(4) Although I cannot provide the actual image and bounding box, I can tell you the names of the entities specified by the bounding box and provide relevant knowledge about these entities to aid in designing the textual question and answer. Absolutely do not disclose the name of the entity in the textual question.
(5) Use only the given knowledge to construct questions; do not make up information, as you cannot actually see the image and confirm its medical modality, therefore avoid using knowledge that might not match the information in the image.
(6) Focus on physiological functions, anatomical positions, biochemical characteristics, pathological features, and diagnostic treatment characteristics from a medical professional perspective. Explore potential relationships among all medical entities within the bounding box from these professional perspectives. Each question must involve all entities specified by the bounding box; questions about a single entity are not acceptable.
(7) In the [textual answer], explain the names of the entities included in the bounding box and provide a systematic and comprehensive answer integrating all the given knowledge.
<Specified medical entities in the bounding box(or point)>: [Meta Data],
<Knowledge>: [Related Knowledge];

Figure 12. Knowledge-based Prompt for Q&A generation on multi-label data in Perspective IV.

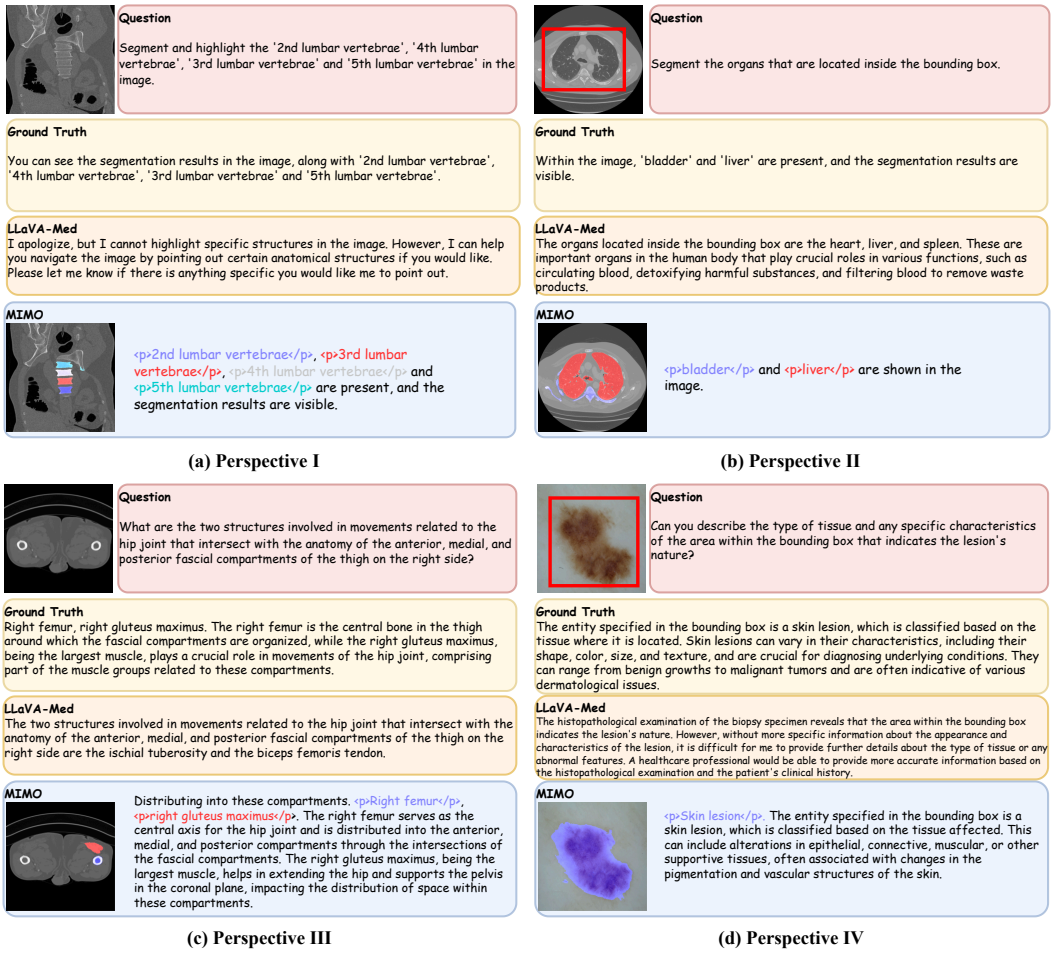

Figure 13. Qualitative analysis compared with LLaVA-Med.

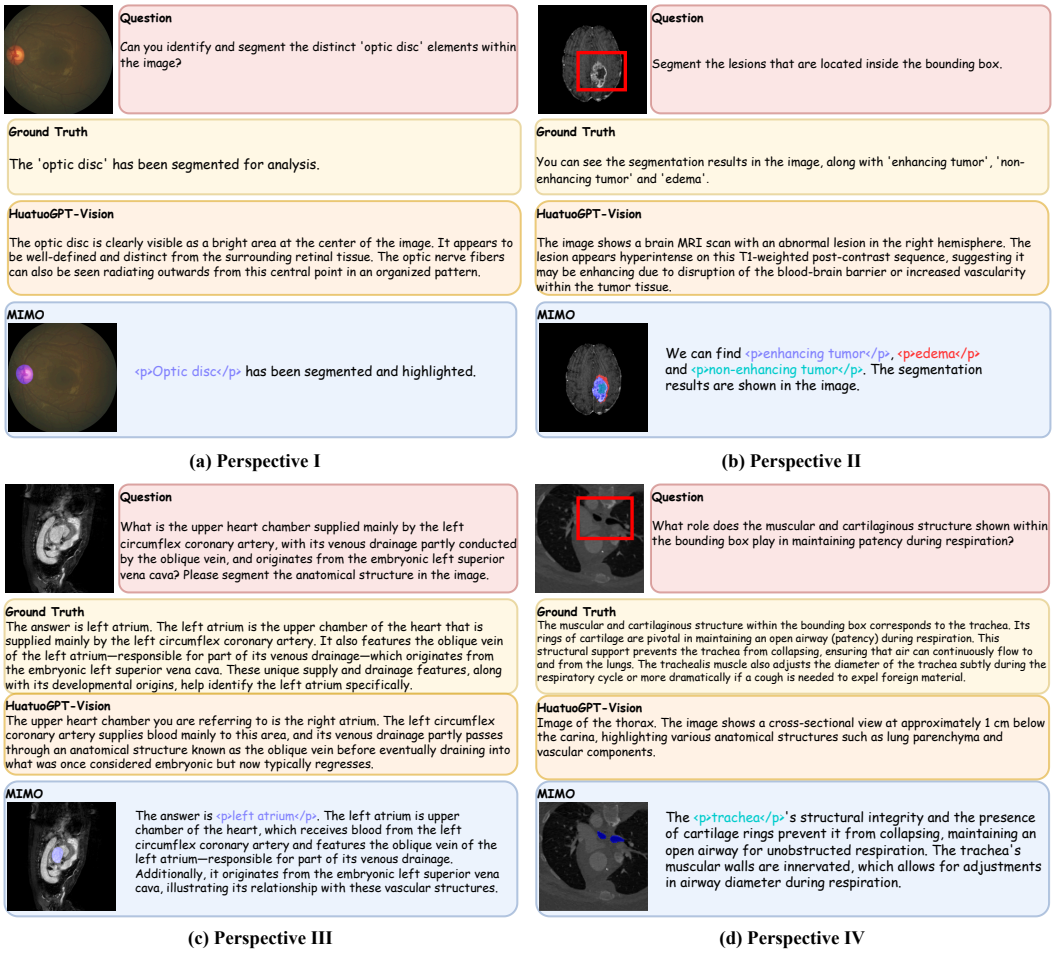

Figure 14. Qualitative analysis compared with HuatuoGPT-Vision.

Table 10. Data source of MIMOSeg. P-I, P-II, P-III, and P-IV respectively represent whether the dataset is applied to the data construction of Perspective I, Perspective II, Perspective III, and Perspective IV.

Datasets	Modality	Labels	Images	Masks	P-I	P-II	P-III	P-IV
CT-ORG[67]	CT	Kidney,Lung,brain,etc.	30000	76171	✔	✔	✘	✘
BraTS2021[7]	MR(FLAIR,T1CE)	Edema,Tumor	59963	117335	✔	✔	✘	✘
BraTS2021[7]	MR(T1,T2)	Edema	58228	58228	✘	✘	✔	✔
EndoVis-2017-RIS[3]	Endoscopy	Shaft, Wrist, Clasper	2978	8530	✔	✔	✘	✘
endovis15[23]	Endoscopy	Polyp	612	612	✘	✘	✔	✔
BrainTumour[16]	MR(T1W,T2W)	Edema,Tumor	59951	118607	✔	✔	✘	✘
BrainTumour[16]	MR-FLAIR	Edema,Tumor	29357	29357	✘	✘	✔	✔
ATM2022[94]	CT	Respiratory Tract	39227	39227	✔	✔	✘	✘
LNDb[60]	CT	Lung Nodule	962	962	✔	✔	✘	✘
VerSe20[73]	CT	Vertebrae	29997	77775	✔	✔	✘	✘
ISLES_SISS[53]	MR(DWI,T1)	Ischemic Stroke	6527	6527	✔	✔	✘	✘
ISLES_SISS[53]	MR(FLAIR,T2)	Ischemic Stroke	6538	6538	✘	✘	✔	✔
ISLES-SPES[53]	MR(T2,DWI)	Ischemic Stroke	6541	6541	✔	✔	✘	✘
ISLES-SPES[53]	MR(cbf,ttp,tmax,cbv,t1c)	Ischemic Stroke	16395	16395	✘	✘	✔	✔
ISLES2018[53]	CT(TMAX,CBV)	Ischemic Stroke	565	565	✔	✔	✘	✘
ISLES2018[53]	CT-MTT	Ischemic Stroke	283	283	✘	✘	✔	✔
ISLES2022[61]	MR-ADC	Ischemic Stroke	5919	5919	✔	✔	✘	✘
ISLES2022[61]	MR-DWI	Ischemic Stroke	5928	5929	✘	✘	✘	✔
Totalsegmentator-dataset[57]	CT	Spleen,Hip,Urinary Bladder,etc.	140624	1055734	✘	✘	✘	✔
autoPET[98]	PET	Tumor	6019	6019	✔	✔	✘	✘
VESSEL2012[70]	CT	Lung Vessel	20405	20405	✔	✔	✘	✘
Brain-PTM[4]	MR-T1	Matter Tracts	11402	11402	✔	✔	✘	✘
AMOS2022[34]	MR	Spleen,Duodenum,Esophagus,etc.	6391	28974	✔	✔	✘	✘
AMOS2022[34]	CT	Spleen,Duodenum,Esophagus,etc.	5985	10169	✘	✘	✘	✔
CTSpine1K-Full[21]	CT	Vertebrae	29998	82162	✔	✔	✘	✘
AbdomenCT1K[52]	CT	Pancreas,Kidney,Spleen,Liver	40000	86070	✔	✔	✘	✘
CTPelvic1k[47]	CT	Hip,Sacrum,Lumbar Vertebra	30000	63882	✔	✔	✘	✘
CrossMoDA22[22]	MR-T1CE	Vestibular Schwannoma	1476	1476	✔	✔	✘	✘
Ultrasound-nerve[55]	Ultrasound	Ultrasound Nerve	2320	2320	✔	✔	✘	✘
Isic2016-task1[29]	Dermoscopy	Skin Lesion	1279	1279	✘	✘	✔	✔
Isic2017-task1[19]	Dermoscopy	Skin Lesion	2746	2746	✔	✔	✘	✘
Isic2018-task1[18]	Dermoscopy	Skin Lesion	2689	2689	✔	✔	✘	✘
LongitudinalMultiple[11]	MR-T2	Multiple Sclerosis Lesion	2025	2025	✔	✔	✘	✘
LongitudinalMultiple[11]	MR-FLAIR	Multiple Sclerosis Lesion	976	976	✘	✘	✔	✔
mnms2[54]	MR	Ventricle,Ventricular Myocardium,etc.	2477	6779	✔	✔	✘	✘
COVID19CTscans[63]	CT	Lung,Lung Infectiions	6719	14255	✔	✔	✘	✘
ASC18[17]	MR-LGE	Left Atrium	8210	8210	✔	✔	✘	✘
cvc-clinicdb[9]	Endoscopy	Polyp	612	612	✔	✔	✘	✘
hvsmr-2016[58]	MR	Heart Blood Pool,Myocardium,etc.	1979	5224	✔	✔	✘	✘
PALM[24]	Fundus-Photography	Optic Disc	1144	1144	✔	✔	✘	✘
COVID-19-20[68]	CT	COVID	4794	4794	✔	✔	✘	✘
Prostate-MRI[48]	MR-T2W	Prostate	1861	1861	✔	✔	✘	✘
MSD-Lung[76]	CT	Lung Cancer	2313	2313	✔	✔	✘	✘
MSD-Spleen[76]	CT	Spleen	1004	1004	✘	✘	✔	✔
MSD-Heart[76]	MR	Left Atrium	1081	1081	✘	✘	✔	✔
MSD-Prostate[76]	MR-ADC	Prostate Zone	365	497	✘	✘	✔	✔
Instance22[75]	CT	Intracranial Hemorrhage	698	698	✔	✔	✘	✘
picai-semi[71]	MR-HBV	Prostate Cancer	414	414	✔	✔	✘	✘
picai-baseline[71]	MR-HBV	Prostate Cancer	314	314	✘	✘	✔	✔
Pulmonary-Chest[33]	X-Ray	Left Lung,Right Lung	36	72	✔	✔	✘	✘
Parse22[49]	CT	Pulmonary Artery	34684	34684	✘	✘	✘	✔
KiTS[31]	CT	Kidney Tumor,Kidney	33112	40874	✘	✘	✔	✔
ACDC[72]	MR	Ventricle,myocardium	1835	3554	✘	✘	✔	✔
LUNA16[74]	CT	Lung	30000	30000	✘	✘	✘	✔
MMWHS[26]	CT	Pulmonary Artery,Atrium Blood Cavity,etc.	2199	3368	✘	✘	✔	✔
FLARE21[51]	CT	Pancreas,Kidney,Liver,Spleen	20000	34454	✘	✘	✘	✔
CAD-PE[27]	CT	Pulmonary Embolism	6667	6667	✘	✘	✘	✔
Chest-Image-Pneum[35]	X-Ray	Pneumothorax	2426	2426	✘	✘	✔	✔
WORD[50]	CT	Femur Head,Stomach,Spleen,Duodenum,etc.	16571	28742	✘	✘	✔	✔
PROMISE12[42]	MR	Prostate	776	776	✘	✘	✔	✔
gamma[83]	Fundus-Photography	Optic Disc,Optic Cup	180	180	✘	✘	✔	✔
cranium[17]	CT	Intracranial Hemorrhage	210	210	✘	✘	✔	✔