MedSG-Bench: A Benchmark for Medical Image Sequences Grounding

Jingkun Yue[1] **Siqi Zhang**[1] **Zinan Jia**[1] **Huihuan Xu**[1]
Zongbo Han[1] **Xiaohong Liu**[2] **Guangyu Wang**[1]*

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications [2]South China Hospital, Medical School, Shenzhen University

Abstract

Visual grounding is essential for precise perception and reasoning in multimodal large language models (MLLMs), especially in medical imaging domains. While existing medical visual grounding benchmarks primarily focus on single-image scenarios, real-world clinical applications often involve sequential images, where accurate lesion localization across different modalities and temporal tracking of disease progression (e.g., pre- vs. post-treatment comparison) require fine-grained cross-image semantic alignment and context-aware reasoning. To remedy the underrepresentation of image sequences in existing medical visual grounding benchmarks, we propose MedSG-Bench, the first benchmark tailored for **Med**ical Image **S**equences **G**rounding. It comprises eight VQA-style tasks, formulated into two paradigms of the grounding tasks, including 1) Image Difference Grounding, which focuses on detecting change regions across images, and 2) Image Consistency Grounding, which emphasizes detection of consistent or shared semantics across sequential images. MedSG-Bench covers 76 public datasets, 10 medical imaging modalities, and a wide spectrum of anatomical structures and diseases, totaling 9,630 question–answer pairs. We benchmark proprietary models (e.g., GPT-4o), general-purpose MLLMs (e.g., Qwen2.5-VL) and medical-domain specialized MLLMs (e.g., HuatuoGPT-vision), observing that even the advanced models exhibit substantial limitations in medical sequential grounding tasks. To advance this field, we construct MedSG-188K, a large-scale instruction-tuning dataset tailored for sequential visual grounding, and further develop MedSeq-Grounder, an MLLM designed to facilitate future research on fine-grained understanding across medical sequential images. We release all resources on `https://github.com/Yuejingkun/MedSG-Bench`

1 Introduction

Visual grounding is the key step that transforms MLLMs from coarse alignment between language expressions and corresponding visual regions to fine-grained visual understanding and reasoning[1]. For example, models like ChatGPT O3[2] often first identify image regions relevant to the questions during reasoning, which helps reduce hallucinations and enhances the trustworthiness of the results. This capability is particularly crucial in medical imaging, where understanding the semantic content of clinical text (e.g., radiology reports) and accurately localizing the corresponding pathological regions is essential for interpretable and reliable diagnosis[3, 4, 5, 6].

Currently, existing medical visual grounding benchmarks focus mainly on single-image scenarios [7, 8]. However, real-world clinical diagnosis inherently requires sequential image analysis. As

*Corresponding author: guangyu.wang24@gmail.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

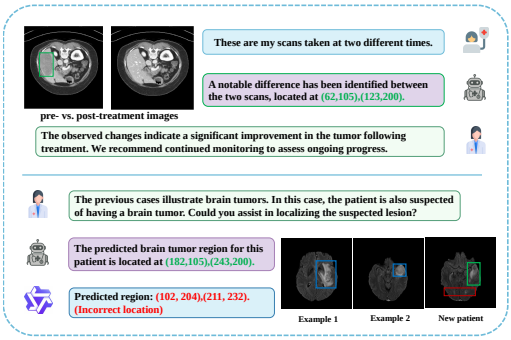

Figure 1: Examples of medical image sequences grounding.

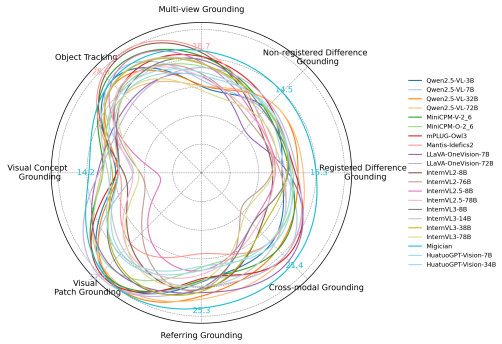

Figure 2: Comparing mainstream MLLMs on MedSG-Bench.

illustrated in Fig. 1, when assessing disease progression, clinicians routinely perform cross-image comparison (pre- vs. post-treatment images), tracking lesion evolution by analyzing changes in size, morphology, and signal intensity across longitudinal CT scans rather than relying solely on a single static image[9]. This essential practice of lesion localization and semantic alignment across multiple images forms the cornerstone of reliable clinical reasoning, yet remains underrepresented in current benchmarks.

To address this gap, we introduce MedSG-Bench, the first comprehensive benchmark specifically designed for medical visual grounding in sequential images. Built upon 76 publicly available medical imaging datasets, covering 10 imaging modalities, and 114 clinical tasks, our benchmark systematically evaluates cross-image grounding capability. Specifically, MedSG-Bench consists of eight carefully designed VQA-style tasks, organized into two grounding paradigms: 1) Image Difference Grounding, which targets the detection of differing regions between sequential images, and 2) Image Consistency Grounding, which focuses on discovering semantically consistent or shared regions across image sequences. This dual-paradigm grounding benchmark can evaluate the essential clinical competencies required for medical image analysis.

In summary, the contributions of this work are as follows:

1. We introduce MedSG-Bench, the first benchmark comprising 9,630 VQA-style samples specifically designed to evaluate the grounding capabilities of MLLMs in medical image sequences. The benchmark defines eight tasks grouped into two core paradigms, Image Difference Grounding and Image Consistency Grounding, which jointly serve to evaluate essential clinical competencies required for medical image analysis.

2. We conduct comprehensive evaluations of proprietary models (e.g., GPT-4o[10]), general-purpose MLLMs (e.g., Qwen2.5-VL[11]) and medical-domain specialized MLLMs (e.g., HuatuoGPT-Vision[12]) on MedSG-Bench. Our results (Fig. 2) show that all current MLLMs exhibit substantial limitations in fine-grained grounding of medical image sequences.

3. To promote progress in this underexplored area, we construct MedSG-188K, a large-scale instruction-tuning dataset tailored for grounding in medical image sequences. Based on this dataset, we further develop MedSeq-Grounder, and achieves state-of-the-art performance on MedSG-Bench.

2 Related work

2.1 Multimodal Large Language Models

Recent advances in multimodal large language models (MLLMs) have progressively extended their capabilities from coarse image-level understanding to fine-grained visual grounding[1, 13]. This progress has been primarily achieved through three main approaches: 1) instruction tuning with grounding supervision[14, 15], 2) integrating external localization modules[16, 17, 18, 19, 20, 21, 22] such as SAM[23] or Grounding DINO[24], and 3) leveraging vision tokenizers to enable perceive-then-understand paradigms[25, 26]. While these methods have significantly improved grounding

Table 1: Comparison between MedSG-Bench and other existing benchmarks in the medical field. FG denotes fine-grained annotation. * indicates the test set.

Benchmark	Size	Task	Multi-modality	Multi-organ	Image-Sequence	FG	Max Length
Understanding-oriented medical benchmarks							
VQA-RAD[33]	3K	11	✓	✓	✗	✗	1
SLAKE*[29]	2K	10	✓	✓	✗	✓	1
OmniMedVQA[34]	128K	5	✓	✓	✗	✗	1
GMAI-MMBench[30]	26K	18	✓	✓	✗	✓	1
Medical-Diff-VQA*[31]	70K	7	✗	✗	✓	✗	2
MMXU*[9]	3K	3	✗	✗	✓	✓	2
Grounding-oriented medical benchmarks							
MS-CXR*[7]	1K	1	✗	✗	✗	✓	1
MeCoVQA-G*[8]	2K	1	✓	✓	✗	✓	1
MedSG-Bench	9K	8	✓	✓	✓	✓	6

accuracy within individual images, they largely overlook the clinically relevant and more complex setting of multi-image visual grounding. MC-Bench[27] first introduced the multi-context visual grounding task and Migician[28] is the first model to tackle this challenge in the natural image domain, enabling free-form and accurate grounding across multiple images. Building upon this paradigm, we extend the exploration to the medical domain, focusing on sequential visual grounding in clinically meaningful scenarios.

2.2 Medical MLLM Benchmarks

As shown in Table 1, benchmarks in the medical domain have progressed from early settings involving single-image and single-modality inputs to more advanced configurations covering multiple organs[29], cross-modal scenarios[30], and multi-image understanding[31, 9]. Some recent benchmarks[32] have also provided fine-grained annotations to enrich evaluation. However, these benchmarks primarily emphasize image-level understanding. Even when detailed annotations are available, they are typically utilized for classification or question answering tasks, rather than for explicit visual grounding. In contrast, grounding-oriented benchmarks remain scarce in the medical domain and are currently limited to single-image scenarios[7, 8]. To date, no medical benchmark has systematically explored sequential visual grounding, a capability that is essential for various clinical tasks such as cross-view lesion comparison, longitudinal disease progression tracking, and multi-phase imaging interpretation. To fill this gap, we propose MedSG-Bench, the first benchmark dedicated to fine-grained visual grounding in sequential medical images.

2.3 Temporal Medical Analysis

Recent studies have increasingly focused on incorporating temporal information to enhance the effectiveness of radiology retrieval and lesion progression detection. Some approaches[35, 36, 37] explicitly integrate temporal data as a feature within the model architecture, allowing the model to directly account for time-based changes in medical images. Other methods[38] treat temporal data as a dynamic semantic signal, improving the retrieval process by enabling the model to capture evolving patterns over time. Both strategies have shown promising results in downstream applications, particularly in medical report generation and disease progression analysis.

3 MedSG-Bench

In this section, we provide an in-depth overview of the careful design and development of MedSG-Bench, covering the rigorous collection and preprocessing of medical data, the systematic definition of tasks tailored for sequential visual grounding, and the presentation of detailed dataset statistics.

3.1 Data Collection and Preprocessing

3.1.1 Dataset Review and Selection

As shown in Fig. 4, open data repositories, including Zenodo, Github, among others, were searched for medical image datasets. Data with permissive licenses (e.g., CC BY 4.0) that allow derivative

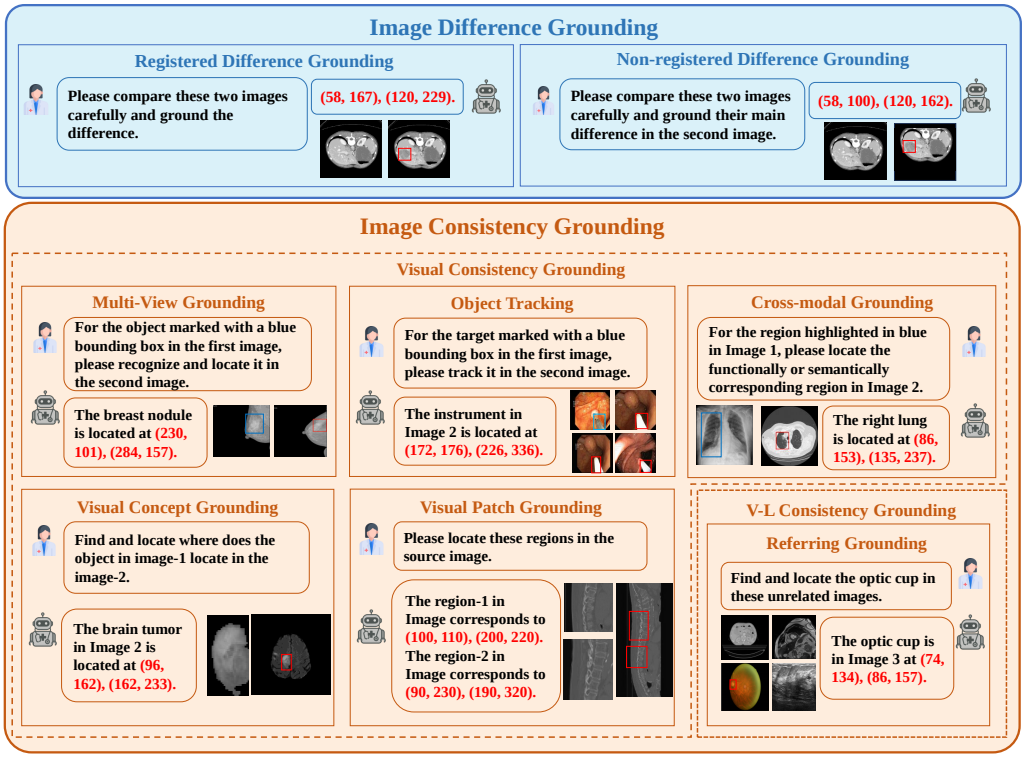

Figure 3: An illustration of medical image sequences grounding tasks included in MedSG-Bench.

works and redistribution were given priority during selection. We retained only those datasets that provided local annotations, such as segmentation masks or bounding boxes, which are essential for grounding-based tasks. To ensure mutual exclusivity among imaging cases, we cross-referenced dataset metadata and associated papers to identify and remove duplicated samples. Additionally, we performed a manual quality review to exclude images with poor visual clarity or unreliable annotations, and verified that all PHI (Protected Health Information) had been properly de-identified in the source datasets, thereby preserving the overall integrity and usability of the data.

3.1.2 Standardization

Medical imaging datasets exhibit high heterogeneity in format, resolution, intensity distribution, and metadata quality, with modality-specific characteristics that differ markedly from natural images. To mitigate this variability, we followed the preprocessing strategy proposed in [39], applying min-max normalization to rescale pixel intensities to a standardized range, thereby enabling more consistent downstream processing. To unify the data format, both 3D volumetric scans and video sequences were converted into 2D RGB images—achieved by slicing along anatomical axes or sampling frames at fixed intervals, respectively. All images were subsequently resized to 336×336 pixels, and each image was assigned a unique identifier encoding its imaging modality and associated task. Finally, all processed images were stored in lossless PNG format to preserve visual fidelity.

3.2 VQA tasks definition and generation

To facilitate fine-grained evaluation of visual grounding for sequential medical images, we define eight VQA-style tasks, organized into two complementary categories, including Image Difference Grounding and Image Consistency Grounding, which collectively capture both semantic changes and invariant features across image sequences, as illustrated in Fig. 3.

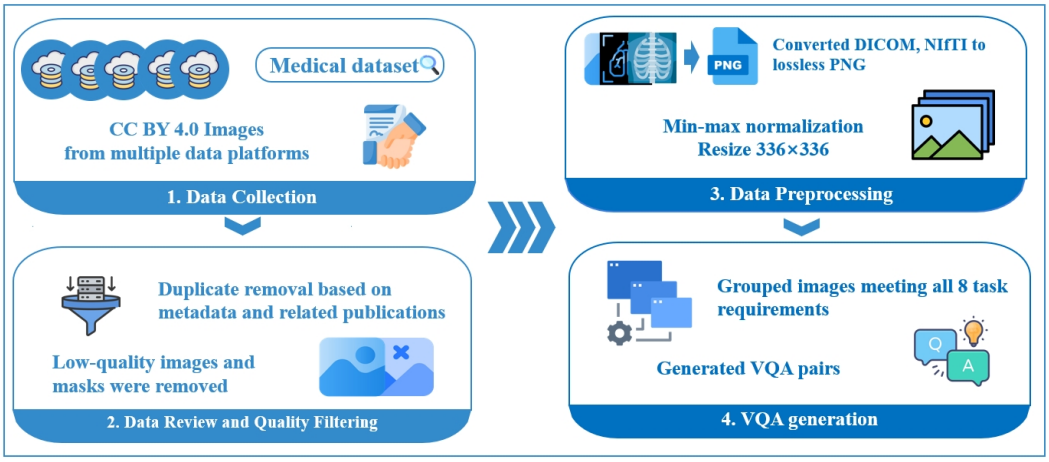

Figure 4: Overview of the MedSG-Bench construction protocol.

3.2.1 Image Difference Grounding

Image Difference Grounding focuses on detecting and localizing regions of changes across sequential images, enabling assessment of a model's ability to perceive subtle or clinically relevant variations.

Task 1: Registered Difference Grounding Given a pair of spatially aligned (i.e., registered) images that are visually identical except for a single region, the model is designed to detect and localize the difference. To generate such image pairs in a controlled and scalable manner, we begin with a single medical image and introduce localized perturbations that simulate clinically meaningful variations, such as disease progression or treatment response. These perturbations comprise both geometric or appearance-based transformations (e.g., CutPaste[40]), and synthetic anomalies generated using state-of-the-art medical generative models[41, 42, 43]. To avoid the model learning shortcuts, such as associating a fixed image position with abnormalities, we randomize the ordering of image pairs, ensuring that either the normal or the abnormal image may appear in either position.

Task 2: Non-registered Difference Grounding In clinical practice, medical images often exhibit spatial misalignments due to patient movement, scanner variability, or imperfect registration. This issue is particularly common when comparing medical images acquired from the same patient at different time points, where the lack of proper registration can lead to spatial shifts in organs or lesions, thereby potentially challenging models to distinguish real differences from registration artifacts. To better simulate such conditions and evaluate the model's robustness to Non-registered Difference Grounding, we extend Task 1 by introducing controlled spatial shifts: each image is randomly translated by up to 20 pixels along both the horizontal and vertical axes. The model is thus required to identify and accurately localize the primary difference between the two images while ignoring changes caused by misalignment.

3.2.2 Image Consistency Grounding

Image Consistency Grounding focuses on identifying and aligning invariant semantics across sequential medical images, which is essential for cross-view, cross-modal and cross-time alignment in clinical practice. Specifically, Image Consistency Grounding can be divided into two subcategories: 1) Visual Consistency Grounding (Task 3-7), which evaluates the model's ability to capture visual consistency across multiple images; 2) Vision-Language Consistency Grounding (Task 8), which involves aligning language-referenced information with multiple medical images.

Task 3: Multi-View Grounding Medical images from different views often have geometric inconsistencies due to patient movement, scanning protocols, or anatomical deformation. To assess a model's ability to capture cross-view correspondence, we construct the Multi-View Grounding task using two implementation strategies. First, we repurpose existing multi-view datasets (e.g., VinDr-Mammo) by converting them into a VQA-style format. Second, we simulate multi-view

Table 2: Detailed statistics of MedSG-Bench.

Task	#Datasets	#Modalities	#Clinical Tasks	Max Length
Registered Difference Grounding	50	10	59	2
Non-registered Difference Grounding	50	10	58	2
Multi-view Grounding	30	4	75	3
Object Tracking	30	4	87	6
Visual Concept Grounding	49	10	87	2
Visual Patch Grounding	53	10	78	5
Cross-modal Grounding	24	4	28	4
Referring Grounding	9	8	28	3
MedSG-Bench	76	10	114	6

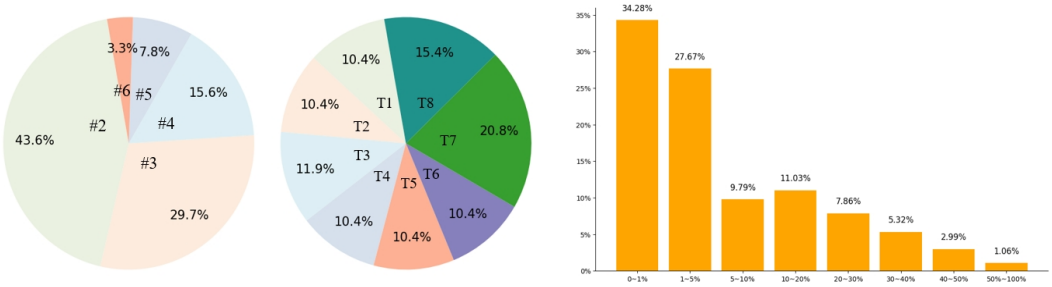

Figure 5: Proportions of image sequence length (**left**), data distribution across tasks (**middle**), and target-to-image size ratios (**right**) in MedSG-Bench.

scenarios by extracting three orthogonal slices (axial, sagittal, and coronal) from 3D medical volumes. Notably, the reference view is not fixed and may vary across different samples.

Task 4: Object Tracking Accurately tracking anatomical structures or instruments across slices of medical images or frames of surgical video is essential in clinical workflows (e.g., lesion monitoring and intraoperative navigation). This task evaluates the model's ability to maintain consistent localization of a target object across sequential frames or slices. We construct this task using two types of data sources. First, we leverage existing surgical videos, where objects such as instruments or tissues are manually annotated across frames. Second, we simulate spatial tracking scenarios by slicing 3D medical volumes along a fixed anatomical axis, treating anatomical structures or lesions as trackable targets across ordered 2D slices.

Task 5: Visual Concept Grounding In clinical scenarios, lesions can exhibit high variability in locations (e.g., across anatomical regions) and visual appearance due to imaging protocols or disease subtypes. This variability challenges models to learn robust target representations based on pathological features, rather than over-relying on spatial biases. This task evaluates the model's ability to recognize and localize a visually distinct and semantically coherent concept, including both pathological findings such as tumors and anatomical structures such as organs or tissue subtypes, within a complex medical image. The model is provided with a reference image in which the concept appears under idealized conditions, and must identify the corresponding instance in a target image with greater visual clutter and contextual complexity. To construct this task, the reference concept is extracted from the target using segmentation masks to ensure semantic consistency.

Task 6: Visual Patch Grounding Precisely distinguishing nearly identical anatomical structures (e.g., separating tumor margins from adjacent vasculature) is essential for image-guided interventions and radiotherapy planning, where subtle visual distinctions determine procedural success. Therefore, we design this task evaluates the model's ability to match a local image patch to its original location within a larger image. It poses significant challenges in contexts where structures like vertebral segments (e.g., T1 to T12) exhibit nearly identical appearances. To construct this task, we initially sample 15 patches per image and manually select up to five based on foreground richness, including organ boundaries, lesion areas, or diagnostically relevant fine structures. The rest are discarded. This

selective sampling ensures that each retained patch presents a non-trivial grounding challenge while avoiding visually homogeneous regions.

Task 7: Cross-modal Grounding In clinical practice, the same patient is often examined using different imaging modalities such as CT, X-ray, or MRI, each highlighting distinct but complementary aspects of anatomical structures or pathologies. This task assesses the model's ability to ground semantically or functionally equivalent regions across differing imaging contexts. Given a reference region from one image, the model is required to identify the corresponding region in a target image that may differ in imaging modality (e.g., CT versus MRI) or contrast type (e.g., T1-weighted versus T2-weighted MRI). Region pairs are manually curated based on metadata such as modality type and annotated labels to ensure semantic alignment and multimodal consistency.

Task 8: Referring Grounding Clinicians often describe findings or refer to specific regions using natural language expressions. Enabling models to accurately interpret and associate such expressions with visual content is essential for enhancing interpretability, supporting human-AI collaboration, and building reliable decision support systems. Considering the prevalence of partially labeled data in medical imaging, we carefully curate candidate image sets to ensure that the images are semantically unrelated. This reduces the risk of referential ambiguity caused by overlapping content or latent correlations among images.

3.3 Data description

We curated a total of 76 publicly available datasets under permissive licenses, prioritizing those released with open CC-BY terms to ensure broad accessibility. As summarized in Table 2, MedSG-Bench spans 10 medical imaging modalities (CBCT, CT, CTA, Colonoscopy, Dermoscopy, Endoscopy, Fundus, MRI, US, X-ray) and and encompasses 114 distinct clinical tasks, covering a wide range of anatomical regions and disease types. The benchmark contains 9,630 visual question answering pairs, derived from 24,341 medical images, designed to assess fine-grained grounding capabilities across diverse clinical contexts. In addition to task coverage, we also provide detailed statistics on the proportion of image sequence lengths, data distribution, and target-to-image size ratios (lesions or anatomical abnormalities are often subtle, localized, and small in size), offering a comprehensive overview of the benchmark's complexity and representativeness in Fig. 5.

4 MedSG-188K and MedSeq-Grounder

4.1 MedSG-188K

The construction of MedSG-188K is based on the eight tasks defined by MedSG-Bench. To ensure diversity in VQA-style queries, we first crafted seed instruction templates tailored to the specific characteristics of each task, capturing the nuanced demands of distinct clinical scenarios. To mitigate potential bias and enhance linguistic diversity, we employed multiple large language models (LLMs), including GPT-4[44], Claude[45], and DeepSeek[46], to expand the seed instruction templates. These models collectively generated ten diverse free-form instruction variants per task by systematically varying the phrasing, contextual framing, and query structure. For each medical image sequence, one of the instruction templates was randomly selected and populated with task-specific content to generate diverse question-answer pairs. Using this pipeline, we constructed a total of 188,163 VQA-style samples, derived from 324,359 medical images. The distribution of sequence lengths, data volume is summarized in Fig. 6.

4.2 MedSeq-Grounder

MedSeq-Grounder is developed based on the Qwen2.5-VL-7B model[11] and trained using the LLaMA-Factory framework[47]. The training is performed with a global batch size of 64 over 15,000 steps, using a learning rate of 5e-6 and 4×A40-48G GPUs.

Table 3: Performance of different MLLMs on MedSG-Bench. IDG: Image Difference Grounding; ICG: Image Consistency Grounding; RDG: Registered Difference Grounding; NRDG: Non-registered Difference Grounding; MV: Multi-view Grounding; OT: Object Tracking; VCG: Visual Concept Grounding; VPG: Visual Patch Grounding; CMG: Cross-modal Grounding; RG: Referring Grounding; Avg.: Average; IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	IDG		ICG						Avg.
		RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	
Proprietary MLLMs										
GPT-4o[10]	–	2.42 0.40	3.45 0.20	16.51 8.62	28.19 23.90	13.18 4.70	38.05 26.40	16.02 4.95	23.08 18.02	17.70 10.60
Claude Sonnet 4[45]	–	0.67 0.00	0.81 0.10	12.56 3.57	23.11 16.50	6.93 1.40	27.44 13.80	9.04 1.80	19.57 10.80	12.51 5.76
Gemini 2.5 Pro[48]	–	9.36 3.20	7.29 2.00	14.26 6.71	19.32 13.80	14.94 10.70	41.11 49.20	24.44 28.12	28.12 22.67	20.66 15.61
General-purpose MLLMs										
Qwen2.5-VL[11]	3B	0.59 0.30	1.62 1.30	7.12 3.90	21.32 16.80	6.98 0.80	27.36 3.40	10.02 1.65	12.99 6.82	10.94 4.20
Qwen2.5-VL[11]	7B	0.88 0.30	1.25 0.00	8.48 3.73	22.41 17.80	4.22 1.00	28.87 5.70	16.29 4.45	12.58 6.21	12.31 4.90
Qwen2.5-VL[11]	32B	2.69 1.40	3.48 1.20	7.35 2.61	19.12 13.40	6.53 1.30	26.92 7.10	12.59 4.90	18.71 11.67	12.47 5.71
Qwen2.5-VL[11]	72B	4.37 2.60	3.46 0.80	7.22 2.78	13.11 7.70	10.33 3.50	26.45 6.30	16.32 7.00	20.19 14.10	13.35 6.12
MiniCPM-V-2_6[49]	8B	1.36 0.00	1.50 0.00	15.82 5.20	24.03 18.50	9.90 2.10	28.65 12.20	12.72 3.30	12.44 3.64	13.24 5.27
MiniCPM-O-2_6[50]	8B	1.69 0.10	1.63 0.00	12.11 2.43	15.25 9.60	9.88 1.70	22.96 9.20	9.53 2.35	8.82 2.02	10.12 3.23
mPLUG-Owl3[51]	7B	2.12 0.00	2.55 0.00	15.64 3.64	15.62 4.40	6.80 0.80	30.42 3.60	17.06 4.80	11.92 5.47	13.22 3.19
Mantis-Idefics2[52]	8B	0.49 0.00	0.62 0.00	18.69 8.59	28.04 23.50	6.27 0.50	10.26 1.10	9.59 0.95	6.05 0.54	9.90 3.91
LLaVA-OneVision[53]	7B	1.09 0.00	0.01 0.00	9.26 1.13	10.50 3.20	11.33 1.80	22.20 5.30	19.08 6.70	17.11 5.67	12.39 3.47
LLaVA-OneVision[53]	72B	2.58 0.80	2.87 0.90	11.74 1.39	9.61 2.30	10.95 3.30	32.38 20.30	16.24 5.40	15.43 6.68	13.21 5.18
InternVL3[54]	8B	1.07 0.30	1.20 0.00	14.36 4.42	13.30 6.50	6.43 0.90	18.73 4.60	4.73 1.15	15.16 7.42	9.26 3.19
InternVL3[54]	14B	0.66 0.00	0.71 0.00	13.24 5.31	19.77 13.00	8.60 2.10	13.17 2.40	10.87 3.70	14.57 7.76	10.53 4.41
InternVL3[54]	38B	0.98 0.10	1.76 0.20	12.99 4.79	19.27 13.60	7.63 2.10	17.76 2.90	6.47 1.75	16.59 10.05	10.37 4.44
InternVL3[54]	78B	0.20 0.00	0.53 0.00	6.35 2.43	13.03 8.00	3.57 0.90	11.81 2.50	3.34 0.85	12.76 8.10	6.44 2.90
Migician[28]	7B	15.26 7.80	14.49 6.10	18.16 7.84	21.38 14.90	14.23 7.20	28.87 13.70	21.41 12.15	25.30 18.02	20.29 11.39
Medical-domain specialized MLLMs										
MedGemma[55]	4B	0.45 0.00	0.84 0.00	7.80 4.53	26.82 22.40	11.31 0.90	26.59 15.40	5.92 0.50	10.01 1.01	10.55 4.82
HuatuoGPT-Vision[12]	7B	1.35 0.00	1.84 0.20	10.42 2.78	14.57 9.20	7.99 0.80	15.52 2.30	9.46 2.15	9.60 1.82	8.97 2.36
HuatuoGPT-Vision[12]	34B	1.44 0.00	2.15 0.00	9.41 1.65	13.25 8.30	6.43 0.70	14.53 1.40	10.60 2.60	8.60 1.75	8.57 2.09
MedSeq-Grounder (Ours)	7B	**83.29** **93.20**	**83.72** **94.10**	**55.03** **60.19**	**62.10** **67.20**	**74.11** **82.60**	**85.25** **98.80**	**78.77** **82.75**	**60.43** **65.59**	**72.55** **79.71**

5 Experiments

5.1 Experiment setup

In this study, we evaluate model performance under a zero-shot setting, where the models were prompted to perform inference without access to in-context examples. We use average Intersection over Union (IoU) and ACC@0.5 as the evaluation metric.

5.2 Models

We benchmark a diverse collection of state-of-the-art MLLMs on MedSG-Bench, including 1) proprietary models, 2) general-purpose models that have extended capabilities in the medical domain, and 3) medical-domain specialized models that are meticulously trained for clinical medicine. All models support image sequence input and span parameter scales from approximately 3 billion to 70 billion. For public models, we use publicly released checkpoints from their official Hugging Face repositories[56], selecting the latest or best-performing version within each model family. For proprietary models, we utilize their respective APIs to access the latest available versions.

Proprietary MLLMs We evaluate GPT-4o[10], Claude Sonnet 4[45], and Gemini 2.5 Pro[48].

General-Purpose MLLMs We evaluate Qwen2.5-VL (3B, 7B, 32B, 72B)[11], MiniCPM-V-2_6[49], MiniCPM-O-2_6[50], mPlug-owl3[51], Mantis-Idefics2[52], llava_onevision (7B, 72B)[53], internvl2 (8B, 78B)[57, 58], internvl2_5 (8B, 78B)[59], internvl3 (8B, 14B, 38B, 78B)[54]. For grounding-oriented MLLMs, we evaluate Migician[28], which supports free-form multi-image grounding and has strong instruction-following capability.

Medical-domain specialized MLLMs We evaluate HuatuoGPT-Vision (7B, 34B)[12], which is built on a large-scale and high-quality medical VQA dataset, PubMedVision, as well as other models such as MedGemma (4B)[55], LLaVA Med v1.5 (7B)[60], and BiMediX2 (8B)[61], which are also trained on specialized medical datasets for medical-domain tasks.

In our evaluation process, we carefully considered the potential impact of inconsistencies in coordinate formats across different models. To address this, we consulted the official documentation or papers for each baseline model to determine the expected coordinate format. For instance, the InternVL series models normalize coordinates to the range [0, 1000], while Qwen2.5-VL supports absolute coordinate output, and Gemini 2.5 Pro uses the format [y_min, x_min, y_max, x_max].

5.3 Main Results

Based on the evaluation results presented in Table 3 and 5, we have some findings as follows:

Grounding in medical image sequences is still challenging for all MLLMs Our MedSG-Bench provides a comprehensive multitask challenge, revealing that even the top-performing model Gemini-2.5-pro is limited to the average IoU of 20.66% and Acc@0.5 of 15.61% in zero-shot setting. In particular, most MLLMs struggle with the Image Difference Grounding task. Moreover, the most advanced models do not consistently excel across all tasks, for example, while Gemini-2.5-pro achieves relatively high accuracy on the cross-modal grounding task, its performance on multi-view grounding or object tracking remains notably lower than Mantis and GPT-4o, highlighting the challenge of generalization across diverse grounding scenarios. With instruction tuning on our MedSG-188K dataset, the proposed MedSeq-Grounder achieves state-of-the-art performance across all tasks, demonstrating its effectiveness and robustness in sequential medical visual grounding.

All MLLMs exhibit limitations in detecting small medical targets Small target recognition is a critical challenge in the medical domain, we further categorized the targets into three groups based on their bounding box area ratio: small (0-1%), medium (1-10%), and large (>10%). Table 7 demonstrates that most MLLMs exhibit substantially reduced performance on small targets, underscoring their limitations in precise medical sequential grounding. In contrast, MedSeq-Grounder consistently achieves strong performance across all target sizes, demonstrating its robustness grounding capability in clinically challenging scenarios.

Medical-domain specialized models are often worse than general-purpose models While specialist models are explicitly developed for the medical domain, they often underperform non-specialist open-source models. For example, HuatuoGPT-Vision-7B, lags behind Qwen2.5-VL-7B by 3.34% in average IoU and 2.54% in Acc@0.5 on MedSG-Bench. Notably, it even performs worse than the smaller-sized Qwen2.5-VL-3B model. This performance gap may be attributed to the nature of training data used for domain adaptation. Most existing medical instruction-tuning datasets focus predominantly on image-level understanding tasks, such as classification or report summarization. While HuatuoGPT-Vision is built upon Qwen-VL, its further tuning on understanding-centric medical data appears to have degraded its grounding capability. This reflects a case of catastrophic forgetting, where the model's original ability for spatial alignment is compromised due to continued learning on tasks that lack grounding supervision.

Larger or newer models do not guarantee improved grounding performance Although model scale and recency are commonly associated with improved performance, we find that larger or more recently released models do not necessarily exhibit stronger grounding capabilities in medical image sequences. For instance, InternVL2.5-8B and InternVL3-8B both underperform compared to the earlier InternVL2-8B model, despite architectural updates and increased pretraining. Similarly, MiniCPM-O-2_6 lags behind MiniCPM-V-2_6, highlighting that newer instruction-tuned variants may sacrifice grounding performance in favor of improvements on general-purpose understanding tasks. In some cases, such as with the InternVL family, even the 70B-scale model yields worse results on MedSG-Bench compared to its 8B counterpart, indicating that grounding ability may not scale proportionally with model size. These results suggest that many recent models are primarily optimized for high-level semantic tasks, such as open-ended QA or captioning, and are trained on instruction-tuning datasets that provide little to no supervision for spatial localization or visual grounding. This observation further underscores the importance of dedicated benchmarks like MedSG-Bench, which are specifically designed to evaluate fine-grained grounding and spatial alignment across sequential medical images.

6 Conclusion

This work introduces MedSG-Bench, the first benchmark specifically designed to evaluate the fine-grained visual grounding capabilities of MLLMs in sequential medical images. Through systematic evaluations on eight clinically inspired grounding tasks, we find that all current MLLMs exhibit substantial limitations in medical image sequences grounding. To address these challenges, we construct a grounding instruction-tuning dataset, MedSG-188K, and develop MedSeq-Grounder. We hope our benchmark, dataset, and model will together advance the development of visual grounding in medical image sequences.

7 Limitations and Future Work

While MedSG-Bench is constructed from a wide range of publicly available datasets, it does not include private real-world clinical data such as longitudinal studies, multi-timepoint diagnostics, or follow-up imaging records. This limits its ability to fully capture the temporal complexity and diagnostic continuity inherent in actual clinical workflows. Meanwhile, MedSeq-Grounder is a task-specific model for medical image sequences grounding. Directly fine-tuning may reduce its performance on other tasks such as free-text QA.

In future work, we plan to collaborate with medical institutions to incorporate authentic clinical data, including patient trajectories across multiple visits and imaging sessions, to enhance the benchmark's realism and clinical applicability. And we plan to continue training the model on broader medical instruction data beyond grounding tasks to enhance its general multimodal capabilities.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (grants T2522008, 62272055, 82522048, 62425112 and 624B2100), New Cornerstone Science Foundation through the XPLORER PRIZE, Xiaomi Foundation.

References

[1] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*, 2024.

[2] OpenAI. chatgpto3. `https://openai.com/index/thinking-with-images/`, 2025.

[3] Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. Medrg: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*, 2024.

[4] Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023.

[5] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

[6] Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. Gemex-thinkvg: Towards thinking with visual grounding in medical vqa via reinforcement learning. *arXiv preprint arXiv:2506.17939*, 2025.

[7] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[8] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3779–3787, 2025.

[9] Linjie Mu, Zhongzhen Huang, Shengqian Qin, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. Mmxu: A multi-modal and multi-x-ray understanding dataset for disease progression. *arXiv preprint arXiv:2502.11651*, 2025.

[10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[11] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[14] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[15] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[16] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[17] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.

[18] 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.

[19] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[20] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[21] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[22] Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, et al. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[23] 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.

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[25] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024.

[26] Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024.

[27] Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*, 2024.

[28] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*, 2025.

[29] 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.

[30] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[31] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 12:13, 2023.

[32] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*, 2024.

[33] 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.

[34] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[35] Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. Libra: Leveraging temporal images for biomedical radiology analysis. *arXiv preprint arXiv:2411.19378*, 2024.

[36] Xin Mei, Rui Mao, Xiaoyan Cai, Libin Yang, and Erik Cambria. Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4699–4708, 2024.

[37] Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *arXiv preprint arXiv:2505.03401*, 2025.

[38] Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, et al. Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. *IEEE Transactions on Medical Imaging*, 2025.

[39] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[40] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[41] Linshan Wu, Jiaxin Zhuang, Yanning Zhou, Sunan He, Jiabo Ma, Luyang Luo, Xi Wang, Xuefeng Ni, Xiaoling Zhong, Mingxiang Wu, et al. Freetumor: Large-scale generative tumor synthesis in computed tomography images for improving tumor recognition. *arXiv preprint arXiv:2502.18519*, 2025.

[42] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023.

[43] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11147–11158, 2024.

[44] 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.

[45] Anthropic. claudesonnet4. https://www.anthropic.com/claude/sonnet, 2025.

[46] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[47] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

[48] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[50] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. minicpm-o. https://github.com/OpenBMB/MiniCPM-o, 2025.

[51] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[52] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[53] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[54] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[55] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[56] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

[57] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[58] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[59] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[60] 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. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[61] Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, et al. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *arXiv preprint arXiv:2412.07769*, 2024.

[62] Zhouqiang Jiang. 4c2021. `https://aistudio.baidu.com/datasetdetail/89548`, 2021.

[63] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.

[64] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[65] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

[66] Minghui Zhang, Yangqian Wu, Hanxiao Zhang, Yulei Qin, Hao Zheng, Wen Tang, Corey Arnold, Chenhao Pei, Pengxin Yu, Yang Nan, et al. Multi-site, multi-domain airway tree modeling. *Medical image analysis*, 90:102957, 2023.

[67] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

[68] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022.

[69] Pablo Gómez, Andreas M Kist, Patrick Schlegel, David A Berry, Dinesh K Chhetri, Stephan Dürr, Matthias Echternach, Aaron M Johnson, Stefan Kniesburges, Melda Kunduk, et al. Bagls, a multihospital benchmark for automatic glottis segmentation. *Scientific data*, 7(1):186, 2020.

[70] Msoud Nickparvar. Brain tumor mri dataset. *Kaggle*, 2021.

[71] Itzik Avital, Ilya Nelkenbaum, Galia Tsarfaty, Eli Konen, Nahum Kiryati, and Arnaldo Mayer. Neural segmentation of seeding rois (srois) for pre-surgical brain tractography. *IEEE transactions on medical imaging*, 39(5):1655–1667, 2019.

[72] Ilya Nelkenbaum, Galia Tsarfaty, Nahum Kiryati, Eli Konen, and Arnaldo Mayer. Automatic segmentation of white matter tracts using multiple brain mri sequences. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 368–371. IEEE, 2020.

[73] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[74] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[75] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[76] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

[77] Germán González, Daniel Jimenez-Carretero, Sara Rodríguez-López, Carlos Cano-Espinosa, Miguel Cazorla, Tanya Agarwal, Vinit Agarwal, Nima Tajbakhsh, Michael B Gotway, Jianming Liang, et al. Computer aided detection for pulmonary embolism challenge (cad-pe). *arXiv preprint arXiv:2003.13440*, 2020.

[78] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[79] Bram Van Ginneken, Tobias Heimann, and Martin Styner. 3d segmentation in the clinic: A grand challenge. In *MICCAI workshop on 3D segmentation in the clinic: a grand challenge*, volume 1, pages 7–15, 2007.

[80] Weiwei Cui, Yaqi Wang, Yilong Li, Dan Song, Xingyong Zuo, Jiaojiao Wang, Yifan Zhang, Huiyu Zhou, Bung san Chong, Liaoyuan Zeng, et al. Ctooth+: A large-scale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 64–73. Springer, 2022.

[81] Weiwei Cui, Yaqi Wang, Qianni Zhang, Huiyu Zhou, Dan Song, Xingyong Zuo, Gangyong Jia, and Liaoyuan Zeng. Ctooth: a fully annotated 3d dataset and benchmark for tooth volume segmentation on cone beam computed tomography images. In *International Conference on Intelligent Robotics and Applications*, pages 191–200. Springer, 2022.

[82] Chestimage. `https://tianchi.aliyun.com/dataset/83075`, 2020.

[83] Shuo Wang, Chen Qin, Chengyan Wang, Kang Wang, Haoran Wang, Chen Chen, Cheng Ouyang, Xutong Kuang, Chengliang Dai, Yuanhan Mo, et al. The extreme cardiac mri analysis challenge under respiratory motion (cmrxmotion). *arXiv preprint arXiv:2210.06385*, 2022.

[84] XIE Juanying and ZHANG Kaiyun. Xr-msf-unet: Automatic segmentation model for covid-19 lung ct images. *Journal of Frontiers of Computer Science & Technology*, 16(8), 2022.

[85] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, et al. Covid-19 ct lung and infection segmentation dataset. *(No Title)*, 2020.

[86] Holger R Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical image analysis*, 82:102605, 2022.

[87] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[88] Murtadha Hssayeni, M Croock, A Salman, H Al-khafaji, Z Yahya, and B Ghoraani. Computed tomography images for intracranial hemorrhage detection and segmentation. *Intracranial hemorrhage segmentation using a deep convolutional model. Data*, 5(1):14, 2020.

[89] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020.

[90] Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, et al. Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021.

[91] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[92] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014.

[93] Alain Lalande, Zhihao Chen, Thibaut Pommier, Thomas Decourselle, Abdul Qayyum, Michel Salomon, Dominique Ginhac, Youssef Skandarani, Arnaud Boucher, Khawla Brahim, et al. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. *Medical Image Analysis*, 79:102428, 2022.

[94] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 263–274. Springer, 2021.

[95] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

[96] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[97] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[98] Huazhu Fu, Fei Li, Xu Sun, Xingxing Cao, Jingan Liao, Jose Ignacio Orlando, Xing Tao, Yuexiang Li, Shihao Zhang, Mingkui Tan, et al. Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography. *Medical Image Analysis*, 66:101798, 2020.

[99] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

[100] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics*, 50(3):1917–1927, 2023.

[101] Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 80–88. Springer, 2015.

[102] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015.

[103] Xiangyu Li, Gongning Luo, Kuanquan Wang, Hongyu Wang, Jun Liu, Xinjie Liang, Jie Jiang, Zhenghao Song, Chunyue Zheng, Haokai Chi, et al. The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: The instance challenge. *arXiv preprint arXiv:2301.03281*, 2023.

[104] Xiangyu Li, Gongning Luo, Wei Wang, Kuanquan Wang, Yue Gao, and Shuo Li. Hematoma expansion context guided intracranial hemorrhage segmentation and uncertainty estimation. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1140–1151, 2021.

[105] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge. *IEEE transactions on medical imaging*, 38(9):2219–2230, 2019.

[106] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[107] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[108] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022.

[109] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

[110] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas De Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 218–229. Springer, 2021.

[111] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

[112] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[113] Shangqi Gao, Hangqi Zhou, Yibo Gao, and Xiahai Zhuang. Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Medical Image Analysis*, 89:102889, 2023.

[114] Xinzhe Luo and Xiahai Zhuang. X-metric: An n-dimensional information-theoretic framework for groupwise registration and deep combined computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9206–9224, 2022.

[115] Fuping Wu and Xiahai Zhuang. Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6021–6036, 2022.

[116] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

[117] Shumao Pang, Chunlan Pang, Zhihai Su, Liyan Lin, Lei Zhao, Yangfan Chen, Yujia Zhou, Hai Lu, and Qianjin Feng. Dgmsnet: Spine segmentation for mr image by a detection-guided mixed-supervised segmentation network. *Medical image analysis*, 75:102261, 2022.

[118] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2020.

[119] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[120] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[121] N. Bloch. Nci-isbi. `https://www.cancerimagingarchive.net/analysis-result/isbi-mr-prostate-2013/`, 2015.

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

[123] Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunovic, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. Palm: Pathologic myopia challenge. *(No Title)*, 2019.

[124] Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023.

[125] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

[126] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.

[127] Patrik F Raudaschl, Paolo Zaffino, Gregory C Sharp, Maria Francesca Spadea, Antong Chen, Benoit M Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017.

[128] Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16:749–756, 2021.

[129] Jason Dowling, Jurgen Fripp, Peter Greer, Sébastien Ourselin, and Olivier Salvado. Automatic atlas-based segmentation of the prostate: A miccai 2009 prostate segmentation challenge entry. *Worskshop in Med Image Comput Comput Assist Interv*, 24:17–24, 2009.

[130] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[131] Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022.

[132] Mete Ahishali, Aysen Degerli, Mehmet Yamac, Serkan Kiranyaz, Muhammad EH Chowdhury, Khalid Hameed, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. Advance warning methodologies for covid-19 using chest x-ray images. *Ieee Access*, 9:41052–41065, 2021.

[133] Aysen Degerli, Mete Ahishali, Mehmet Yamac, Serkan Kiranyaz, Muhammad EH Chowdhury, Khalid Hameed, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. Covid-19 infection map generation and detection from chest x-ray images. *Health information science and systems*, 9(1):15, 2021.

[134] Aysen Degerli, Mete Ahishali, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Reliable covid-19 detection using chest x-ray images. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 185–189. IEEE, 2021.

[135] Mehmet Yamac, Mete Ahishali, Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Convolutional sparse support estimator-based covid-19 recognition from x-ray images. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1810–1820, 2021.

[136] Hongwei Bran Li, Fernando Navarro, Ivan Ezhov, Amirhossein Bayat, Dhritiman Das, Florian Kofler, Suprosanna Shit, Diana Waldmannstetter, Johannes C Paetzold, Xiaobin Hu, et al. Qubiq: Uncertainty quantification for biomedical image segmentation challenge. *arXiv preprint arXiv:2405.18435*, 2024.

[137] Fei Li, Diping Song, Han Chen, Jian Xiong, Xingyi Li, Hua Zhong, Guangxian Tang, Sujie Fan, Dennis SC Lam, Weihua Pan, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *NPJ digital medicine*, 3(1):123, 2020.

[138] Shishuai Hu, Zehui Liao, and Yong Xia. Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 650–659. Springer, 2022.

[139] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011.

[140] Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical image analysis*, 101:103447, 2025.

[141] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020.

[142] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. *Mohannad ParasLakhani Hussain*, 2019.

[143] Tobias Heimann, Bryan J Morrison, Martin A Styner, Marc Niethammer, and Simon Warfield. Segmentation of knee images: a grand challenge. In *Proc. MICCAI Workshop on Medical Image Analysis for the Clinic*, volume 1. Beijing, China, 2010.

[144] Rashed Karim, Lauren-Emma Blake, Jiro Inoue, Qian Tao, Shuman Jia, R James Housden, Pranav Bhagirath, Jean-Luc Duval, Marta Varela, Jonathan M Behar, et al. Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source ct and mri image database. *Medical image analysis*, 50:36–53, 2018.

[145] Zeyang Yao, Wen Xie, Jiawei Zhang, Yuhao Dong, Hailong Qiu, Haiyun Yuan, Qianjun Jia, Tianchen Wang, Yiyi Shi, Jian Zhuang, et al. Imagetbad: A 3d computed tomography angiography image dataset for automatic segmentation of type-b aortic dissection. *Frontiers in Physiology*, 12:732711, 2021.

[146] HM Gireesha and S Nanda. Thyroid nodule segmentation and classification in ultrasound images. *International Journal of Engineering Research and Technology*, 2014.

[147] Rina D Rudyanto, Sjoerd Kerkstra, Eva M Van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, Ilkay Öksüz, et al. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study. *Medical image analysis*, 18(7):1217–1232, 2014.

[148] Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

[149] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[150] Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

[151] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

[152] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have summarized our study in Section Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We described the limitations in Section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All models evaluated in our experiments are publicly available. In addition, we release the evaluation code, model weights, and dataset to facilitate reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All resources are available at `https://anonymous.4open.science/r/test-ABC123`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The evaluation is in the zero-shot setting. The hyperparameters for training MedSeq-Grounder are detailed in Section 4.2. The data statistics are provided in Section 3.3 and C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This is a dataset paper, we did not propose a new method and did not need to show the statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For evaluations, the models in this paper are all public and only the inference is needed. The computer resources for inferencing these models are well known. For model training, we use 4×A40-48G GPUs and execute eight days.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We collect all datasets from the online platforms with the user's informed consent and used under specific data use agreements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There is no high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original paper or attached the link to the existing assets used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: For the newly proposed datasets, we provide detailed description.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing in this study

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We used GPT-4 to assist in the expansion and refinement of instruction templates, which are central to our dataset and instruction tuning process. Furthermore, we employed existing multimodal large language models (MLLMs) as the foundation for instruction tuning, which constitutes a core methodological component of our approach.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Dataset Details

In this section, we provide the detailed datasets used in MedSG-Bench, including the name of the dataset, the modality, the dimension of data, and the accessible links. As shown in Table 4, MedSG-Bench is constructed from 76 datasets across 10 medical image modalities.

Table 4: Detailed datasets information in MedSG-Bench.

Dataset	Modality	Dim	Accessible links
4C2021[62]	CT	3D	https://aistudio.baidu.com/datasetdetail/89548
AbdomenCT1K[63]	CT	3D	https://github.com/JunMa11/AbdomenCT-1K
ACDC[64]	MRI	3D	https://humanheart-project.creatis.insa-lyon.fr/database/
AMOS22[65]	CT, MRI	3D	https://amos22.grand-challenge.org/
ATM22[66]	CT	3D	https://atm22.grand-challenge.org/
Atria Segmentation[67]	MRI	3D	https://www.cardiacatlas.org/atriaseg2018-challenge/atria-seg-data/
AutoLaparo[68]	Colonoscopy	2D	https://autolaparo.github.io/
BAGLS[69]	Endoscopy	2D	https://www.kaggle.com/datasets/gomezp/benchmark-for-automatic-glottis-segmentation
BraimMRI[70]	MRI	3D	https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
BrainPTM[71][72]	MRI	3D	https://brainptm-2021.grand-challenge.org/
BraTS2020[73][74][75]	MRI	3D	https://service.tib.eu/ldmservice/dataset/brats2020
BUSI[76]	US	2D	https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
CAD-PE[77]	CT	3D	https://ieee-dataport.org/open-access/cad-pe
CAMUS[78]	US	2D	https://www.creatis.insa-lyon.fr/Challenge/camus/
Cause07[79]	MRI	3D	https://cause07.grand-challenge.org/
CBCT3D[80][81]	CBCT	3D	https://toothfairy.grand-challenge.org/
Chestimage[82]	X-Ray	2D	https://tianchi.aliyun.com/dataset/83075
CMRxMotions[83]	MRI	3D	https://www.synapse.org/Synapse:syn28503327/
COVID-19[84]	CT	3D	https://medicalsegmentation.com/covid19/
COVID19CTscans[85]	CT	3D	https://zenodo.org/records/3757476
COVID-19-20[86]	CT	3D	https://covid-segmentation.grand-challenge.org/
Covid19cxr[87]	X-ray	2D	https://github.com/ieee8023/covid-chestxray-dataset
Cranium[88]	CT	3D	https://tianchi.aliyun.com/dataset/82967
CT-ORG[89]	CT	3D	https://www.cancerimagingarchive.net/collection/ct-org/
CTSpine1K[90]	CT	3D	https://github.com/MIRACLE-Center/CTSpine1K
CVC-ClinicDB[91]	Colonoscopy	2D	https://polyp.grand-challenge.org/CVCClinicDB/
DRISHTI-GS[92]	Fundus	2D	https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation
EMIDEC[93]	MRI	3D	https://emidec.com/dataset
EndoTect2020[94]	Colonoscopy	2D	https://osf.io/mh9sj/
EndoVis15[95]	Colonoscopy	2D	https://endovis.grand-challenge.org/
EndoVis2017[96]	Colonoscopy	2D	https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/
GAMMA[97][98][99]	Fundus	2D	https://gamma.grand-challenge.org/Home/
HaN-Seg[100]	CT, MRI	3D	https://zenodo.org/records/7442914
Hvsmr2016[101]	MRI	3D	http://segchd.csail.mit.edu/data.html
I2CVB[102]	MRI	3D	https://i2cvb.github.io/
InSTANCE2022[103][104]	CT	3D	https://instance.grand-challenge.org/
iseg2017[105]	MRI	3D	https://iseg2017.web.unc.edu/download/
ISIC2018[106][107]	Dermoscopy	2D	https://challenge.isic-archive.com/data/#2018
ISLES-ATLAS[108]	MRI	3D	https://atlas.grand-challenge.org/
ISLES-MM[108]	MRI	3D	https://isles22.grand-challenge.org/
JSRT[109]	X-ray	2D	http://imgcom.jsrt.or.jp/minijsrtdb/
KvasirInstrument[110]	Colonoscopy	2D	https://datasets.simula.no/kvasir-instrument/

LMSLS[111]	MRI	3D	https://smart-stats-tools.org/lesion-challenge-2015
LUNA16[112]	CT	3D	https://luna16.grand-challenge.org/Download/
MMWHS[113][114][115][116]	CT, MRI	3D	https://www.ub.edu/mnms/
MRSpineSeg[117][118]	MRI	3D	https://mosmed.ai/datasets/covid19_1110
MSD02[119]	MRI	3D	http://medicaldecathlon.com/
MSD04[120]	MRI	3D	http://medicaldecathlon.com/
MSD05[120]	MRI	3D	http://medicaldecathlon.com/
MyoPS2020[113][114][116]	MRI	3D	https://zmiclab.github.io/zxh/0/myops20/
NCI-ISBI2013[121]	MRI	3D	https://www.cancerimagingarchive.net/analysis-result/isbi-mr-prostate-2013/
PadChest[122]	X-ray	2D	https://bimcv.cipf.es/bimcv-projects/padchest/
PALM[123]	Fundus	2D	https://ieee-dataport.org/documents/palm-pathologic-myopia-challenge
Parse2022[124]	CT	3D	https://parse2022.grand-challenge.org/Dataset/
PCXA[125][126]	X-ray	2D	https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html
PDDCA[127]	CT	3D	https://www.imagenglab.com/newsite/pddca/
Pelvic1K[128]	CT	3D	https://zenodo.org/record/4588403
Promise09[129]	MRI	3D	https://www.na-mic.org/wiki/Training_Data_Prostate_Segmentation_Challenge_MICCAI09
PROMISE12[130]	MRI	3D	https://zenodo.org/records/8026660
QaTa-COV19[131][132][133][134][135]	X-ray	2D	https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset
QUBIQ2020[136]	CT	2D	https://qubiq.grand-challenge.org/
REFUGE[99][137]	Fundus	2D	https://refuge.grand-challenge.org/
RIGA+[138]	Fundus	2D	https://zenodo.org/records/6325549
RIM_ONE[139]	Fundus	2D	https://github.com/miag-ull/rim-one-dl
SegRap2023[140]	CT	2D	https://segrap2023.grand-challenge.org/dataset/
SegTHOR[141]	CT	3D	https://competitions.codalab.org/competitions/21145
SIIM-ACR[142]	X-ray	2D	https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation
SKI10[143]	MRI	3D	https://ski10.grand-challenge.org/
SLAWT[144]	MRI	3D	http://stacom.cardiacatlas.org/
TBAD[145]	CTA	3D	https://www.kaggle.com/datasets/xiaoweixumedicalai/imagetbad
TN-SCUI[146]	US	2D	https://tn-scui2020.grand-challenge.org/
VESSEL12[147]	CT	3D	https://vessel12.grand-challenge.org/
VINDR-Mammo[148]	X-ray	2D	https://www.physionet.org/content/vindr-mammo/1.0.0/
Verse19[149][150]	CT	3D	https://github.com/anjany/verse
WMH[151]	MRI	3D	https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/AECRSD
WORD[152]	CT	3D	https://github.com/HiLab-git/WORD

B Template for Instruction Data Generation

Template for Instruction Data Generation

Task1
"Please examine these two images and provide the coordinates of the area where they differ."
"Compare both images closely and share the coordinates of the discrepancy."
"Look at these two images and tell me the coordinates of the difference between them."
"Carefully analyze these images and provide the coordinates of their difference."
"Examine the two images and give me the coordinates of the region where they differ."
"Can you find the differences between these two images and give me the coordinates?"
"Please inspect these two images and indicate the coordinates of their difference."
"Compare the two images and identify the coordinates of the difference."
"Look closely at the two images and provide the coordinates where they differ."
"Analyze both images and provide the coordinates of the difference between them."
Task2
"Compare these two images carefully and give me the coordinates of their real difference in the second image. Find it and locate it in the second image."
"Please examine both images and identify the real difference that appears in the second one. Provide the coordinates of that difference."
"Carefully analyze the two images. What is the actual visual change in the second image? Mark its coordinates precisely."
"Spot the true difference in the second image when compared with the first. Return the bounding box of that change."
"Look at the two images side by side. What is the meaningful change introduced in the second image? Output its location."
"Your task is to detect the actual difference in the second image compared to the first and report its position in coordinates."
"Inspect the two images and tell me where the real change is in the second one. Output the coordinates of the difference."
"Between the two images, find the true variation that exists in the second image. Return its location in bounding box format."
"Compare the pair of images. Where is the real and only difference in the second image? Provide the coordinates."
"Analyze the difference between these images. Identify and locate the actual modified region in the second image only."
Task3
"The object marked with a red bounding box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is shared by these two images. Locate and identify it in the num image."
"In the first image, the object highlighted with a red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is common to both images. Please recognize and locate it in the num image."
"The object outlined by a red bounding box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) appears in both images. Can you identify and find its position in the num image?"
"The object with a red bounding box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is shared between these two images. Locate and recognize it in the num image."
"The object marked in red in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is common across both images. Find and identify it in the num image."
"The object highlighted by the red box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is shared with the second image. Locate it in the num image and provide its position."
"Both images contain a common object marked with a red bounding box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Find and identify this object in the num image."
"In the first image, the object marked by the red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) appears in both. Can you locate it in the num image?"
"The object in the first image, marked by a red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>), is also in the second image. Identify and locate it in the num image."
"In the first image, the object enclosed by the red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) is the same as in the second image. Locate it in the num image and identify its position."

Template for Instruction Data Generation

Task4

"In the first image, a red bounding box marks a specific object (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Your task is to identify and localize the same object in the num image."
"The object enclosed in red in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) also appears in the num image. Detect and locate it accordingly."
"Focus on the object highlighted by the red box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Find and mark this same object in the num image."
"Observe the red-boxed object in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Identify where it appears in the num image."
"The first image contains an object inside a red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Detect this same object in the num image."
"An object is annotated with a red box in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Determine where the same object appears in the num image."
"Use the red-bounded object in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>) as a reference. Identify its location in the num image."
"Locate in the num image the object that corresponds to the red-marked region in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>)."
"The first image includes an object shown with a red bounding box (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Recognize and localize this same object in the num image."
"Refer to the red-outlined region in the first image (<|box_start|> (x_min, y_min), (x_max, y_max) <|box_end|>). Locate the corresponding object in the num image."

Task5

"Given image-1 and image-2, identify and localize the object from image-1 within image-2."
"Based on the object shown in image-1, determine its corresponding location in image-2."
"Observe the object in image-1. Where does it appear in image-2? Mark the location."
"Find the region in image-2 that corresponds to the object highlighted in image-1."
"Refer to image-1 and locate the same object in image-2."
"Your task is to recognize the object from image-1 and indicate where it is in image-2."
"Using image-1 as a reference, identify the location of the same object in image-2."
"Locate the counterpart of the object shown in image-1 within image-2."
"Match the object in image-1 to its corresponding region in image-2 and provide its location."
"Analyze the object in image-1 and find its equivalent presence in image-2 by marking its location."

Task6

"You are given a source image and several cropped regions. Identify where the num region belongs in the source image."
"Observe the original image and its cropped parts. Locate the num region in the source image."
"Given one complete image and multiple region crops, find where the num one fits in the original image."
"You are shown a source image and some regional cutouts. Point out where the num region comes from."
"Refer to the original image and determine the location of the num region shown afterward."
"Analyze the full image and match the num region image to its location within it."
"Based on the source image, indicate where the num region patch belongs."
"Here is a source image followed by cropped regions. Find the position of the num region in the source."
"You are given a full image and several region patches. Locate the num patch within the source image."

Template for Instruction Data Generation

Task7

"You are given total images. Based on the red bounding box in the first image, locate the corresponding region in the num image that shares a similar function or meaning."

"Among the total provided images, examine the red-highlighted area in the first image and identify the region in the num image that matches it semantically or functionally."

"You are given total images. Consider the red-marked region in the first image. In the num image, find the area that best aligns with it in terms of purpose or meaning."

"From the total images below, determine which region in the num image corresponds to the red-boxed area in the first image."

"You are given total images. Study the red region in the first image. Then, in the num image, identify the location that serves a similar role or conveys a similar idea."

"You are given total images. Take a close look at the red-bounded area in the first image. Locate the corresponding region in the num image that reflects the same concept."

"You are given total images. Focus on the red box in the first image. Your task is to find the equivalent region in the num image that shares its function or meaning."

"You are given total images. Analyze the highlighted region in the first image. In the num image, point out the area that represents the same functional or semantic content."

"Given total images, compare the red-boxed area in the first image with the num image and find the corresponding part."

"You are given total images. Observe the first image where a red region is marked. Identify the most similar region in the num image in terms of functionality or semantics."

Task8

"Identify the bounding box of the region described by the following expression: <|object_ref_start|> object name <|object_ref_end|>."

"Locate the region corresponding to the following structure and provide its bounding box:<|object_ref_start|> object name <|object_ref_end|>."

"What is the bounding box for the region denoted by <|object_ref_start|> object name <|object_ref_end|>?"

"Provide the bounding box for the following entity mentioned in the image: <|object_ref_start|> object name <|object_ref_end|>."

"Identify and annotate the bounding box of <|object_ref_start|> object name <|object_ref_end|>."

"Indicate the bounding box of the area that corresponds to <|object_ref_start|> object name <|object_ref_end|>."

"Determine the coordinates of the bounding box for the target structure: <|object_ref_start|> object name <|object_ref_end|>."

"What is the bounding box for the region denoted by <|object_ref_start|> object name1 <|object_ref_end|> and <|object_ref_start|> object name2 <|object_ref_end|>?"

C Data statistics of MedSG-188K

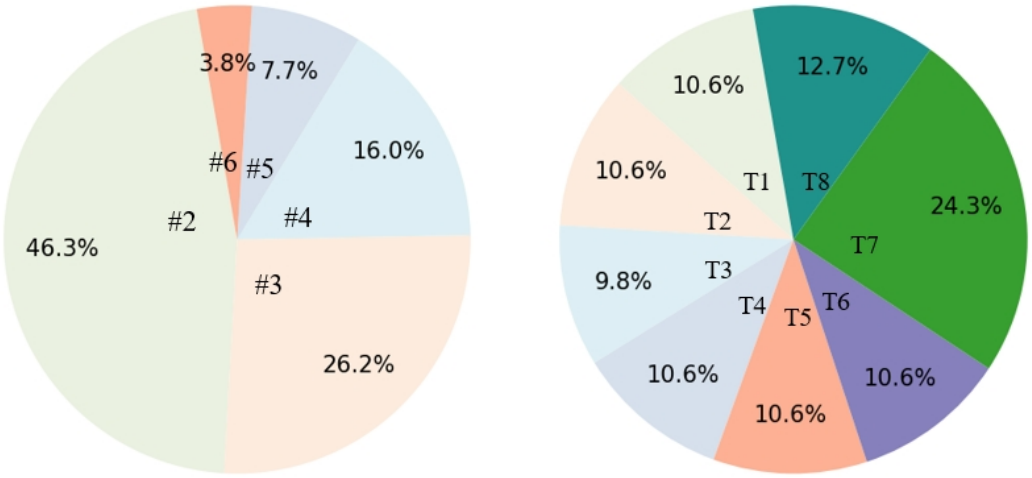

Figure 6: Proportions of image sequence length (**left**), data distribution across tasks (**right**) in MedSG-188K.

D Evaluation Metric

We evaluate model performance using two standard metrics: Intersection over Union (IoU) and Accuracy at IoU threshold 0.5 (Acc@0.5). These metrics are widely adopted in visual grounding to measure localization quality.

IoU quantifies the overlap between the predicted bounding box B_{pred} and the ground-truth bounding box B_{gt}, and is defined as:

$$\text{IoU} = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}})} \tag{1}$$

Acc@0.5 measures the proportion of predictions whose IoU with the ground truth exceeds 0.5. It is defined as:

$$\text{Acc@0.5} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{IoU}_i \geq 0.5) \tag{2}$$

Here, N is the total number of samples, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true, and 0 otherwise.

E Additional Analysis

E.1 More Results

We benchmarked more MLLMs on MedSG-Bench, the results are summarized in Table 5. We grouped the targets by their bounding box area ratio into small (0–1%), medium (1–10%), and large (>10%), and evaluated model performance within each group, the results are summarized in Table 7.

E.2 The potential bias of question generations

To examine whether the use of a single large language model (LLM) introduces bias in our question generation process, we conducted an additional comparative experiment across multiple LLMs, including GPT-4[10], Claude[45], and DeepSeek[46]. All generated questions were manually reviewed to ensure that they accurately preserved the original intent and complied with the standardized instruction format.

We then re-evaluated benchmarked models using questions generated by each LLM individually. The results, summarized in Table 6, report the average performance measured by IoU and acc@0.5. Despite minor variations among the three prompting settings, the observed differences remained within a narrow and acceptable range, suggesting that the core conclusions reported in the Main Results section are stable and robust to the choice of question generator.

Table 5: Performance of other MLLMs on MedSG-Bench. IDG: Image Difference Grounding; ICG: Image Consistency Grounding; RDG: Registered Difference Grounding; NRDG: Non-registered Difference Grounding; MV: Multi-view Grounding; OT: Object Tracking; VCG: Visual Concept Grounding; VPG: Visual Patch Grounding; CMG: Cross-modal Grounding; RG: Referring Grounding; Avg.: Average; IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	IDG		ICG						Avg.
		RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	
General-purpose MLLMs										
InternVL2[58]	8B	0.18 0.00	0.38 0.00	17.34 7.03	26.45 21.20	5.56 0.80	10.36 0.70	6.23 1.00	15.73 7.69	10.24 4.59
InternVL2[58]	76B	0.15 0.00	0.15 0.00	10.00 3.90	15.56 11.80	3.39 0.40	6.64 1.10	2.83 0.75	15.69 9.92	6.88 3.53
InternVL2.5[59]	8B	0.26 0.00	0.38 0.00	13.52 3.56	20.82 13.80	1.96 0.00	5.25 0.00	4.70 0.85	9.56 3.44	7.04 2.56
InternVL2.5[59]	78B	0.24 0.10	0.32 0.10	9.16 2.08	16.18 10.00	4.32 0.50	11.86 2.30	5.48 1.25	10.67 4.52	7.29 2.55
Medical-domain specialized MLLMs										
LLaVA-Med v1.5[60]	7B	0.32 0.00	0.46 0.00	6.45 2.86	11.49 5.61	8.41 0.70	12.74 4.21	6.58 4.35	7.44 1.78	6.29 1.64
BiMediX2[61]	8B	0.24 0.01	0.28 0.00	4.11 1.48	8.66 4.95	7.42 1.12	10.67 3.66	4.38 2.71	7.62 2.02	4.83 1.29

Table 6: Bias Analysis in Question Generation. IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	Avg(GPT-4)	Avg(DeepSeek)	Avg(Claude)	Avg(Ori)
Qwen2.5-VL[11]	3B	10.51 3.86	10.60 3.85	10.31 9.02	10.94 4.20
Qwen2.5-VL[11]	7B	11.25 15.29	10.87 4.13	11.19 4.41	12.31 4.90
Qwen2.5-VL[11]	72B	13.45 6.37	13.35 6.29	13.39 6.41	13.35 6.12
MiniCPM-V-2_6[49]	8B	12.72 4.59	13.33 5.13	12.61 4.30	13.24 5.27
MiniCPM-O-2_6[50]	8B	10.68 3.85	10.34 3.51	10.27 3.32	10.12 3.23
mPLUG-Owl3[51]	7B	10.92 2.86	10.71 2.69	11.04 2.92	13.22 3.19
Mantis-Idefics2[52]	8B	10.33 4.35	10.02 4.07	10.06 3.91	9.90 3.91
LLaVA-OneVision[53]	7B	13.55 5.51	11.59 3.44	12.46 3.47	12.39 3.47
InternVL2.5[59]	8B	7.46 2.78	7.83 2.72	7.13 2.64	7.04 2.56
Migician[28]	7B	20.31 11.39	20.53 11.91	20.43 11.46	20.29 11.39
HuatuoGPT-Vision[12]	7B	9.08 2.71	9.20 2.59	9.18 2.41	8.97 2.36
MedSeq-Grounder (Ours)	7B	72.68 79.98	71.67 78.76	72.86 80.18	72.55 79.71

Table 7: Fine-grained performance of different MLLMs on MedSG-Bench. IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	Avg_small	Avg_medium	Avg_large
General-purpose MLLMs				
Qwen2.5-VL[11]	3B	2.27 1.28	9.53 4.66	24.73 7.47
Qwen2.5-VL[11]	7B	1.69 0.48	8.31 3.74	20.13 7.91
Qwen2.5-VL[11]	32B	3.42 1.21	12.01 5.41	26.46 12.78
Qwen2.5-VL[11]	72B	3.82 1.18	11.96 5.31	25.92 13.25
MiniCPM-V-2_6[49]	8B	2.93 0.55	15.35 6.40	24.93 10.36
MiniCPM-O-2_6[50]	8B	2.60 0.38	10.93 3.25	19.82 7.36
mPLUG-Owl3[51]	7B	2.41 0.00	14.57 22.75	26.86 8.54
Mantis-Idefics2[52]	8B	2.92 1.10	14.47 6.72	12.74 3.51
LLaVA-OneVision[53]	7B	1.19 0.00	15.33 4.28	23.51 7.23
LLaVA-OneVision[53]	72B	3.27 0.55	14.68 3.99	25.38 13.87
InternVL2[58]	8B	2.24 1.01	12.64 6.09	18.09 7.40
InternVL2[58]	76B	1.12 0.32	7.46 3.49	14.39 8.29
InternVL2.5[59]	8B	2.06 0.55	10.24 4.41	9.17 2.54
InternVL2.5[59]	78B	1.33 0.38	8.64 2.86	13.85 5.25
InternVL3[54]	8B	1.81 0.32	9.06 2.49	20.49 8.50
InternVL3[54]	14B	2.23 0.52	12.76 5.12	19.09 8.97
InternVL3[54]	38B	2.47 0.61	11.12 4.70	20.71 9.65
InternVL3[54]	78B	1.26 0.38	6.80 2.57	13.46 7.11
Migician[28]	7B	11.24 4.72	22.05 11.89	30.69 20.36
Medical-domain specialized MLLMs				
HuatuoGPT-Vision[12]	7B	2.55 0.35	11.55 3.88	14.20 2.83
HuatuoGPT-Vision[12]	34B	2.90 0.43	11.02 3.38	12.93 2.41
MedSeq-Grounder (Ours)	7B	68.37 75.83	69.32 75.26	83.88 92.55

E.3 Effect of clinical windowing on model performance

To investigate whether different window settings influence model performance, we conducted additional experiments under two settings: (1) applying only min–max normalization, and (2) applying clinical windowing followed by min–max normalization.

These experiments are conducted on CT datasets, including AbdomenCT1K, LUNA16, and COVID-19-20, where window settings are clinically significant and can substantially affect the visibility of anatomical structures. Specifically, we adopted organ-specific window ranges as follows: Lung: [-600, 1500]; Liver: [60, 150]; Spleen: [60, 150]; Kidney: [40, 400]. The results, summarized in Table 8, report the IoU and acc@0.5 on Registered Difference Grounding, Visual Concept Grounding, and Visual Patch Grounding tasks.

From the results, we observed that clinical windowing led to perfomance gains for models with strong visual perception capabilities (e.g., Migician) or prior exposure to medical data (e.g., HuatuoGPT-Vision), with our proposed MedSeq-Grounder achieving the most significant improvement. We also find that most general-purpose MLLMs typically suffered performance drops under the same setting. This pattern suggests that clinical windowing introduces distribution shifts that challenge general models, while models equipped with robust perceptual abilities and domain-specific knowledge can leverage enhanced contrast and localized visual cues more effectively.

Table 8: Evaluation Results with clinical windowing and min-max normalization on CT datasets. RDG: Registered Difference Grounding; VCG: Visual Concept Grounding; VPG: Visual Patch Grounding; Avg.: Average; IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	RDG		VCG		VPG		Avg	
		ori	window	ori	window	ori	window	ori	window
Qwen2.5-VL[11]	3B	0.31 0.00	0.29 0.00	11.81 1.00	8.87 0.25	28.23 2.93	26.17 1.95	11.67 1.13	10.23 0.64
Qwen2.5-VL[11]	7B	1.15 0.17	0.59 0.00	9.73 1.00	5.98 1.00	26.37 3.41	29.21 4.63	10.90 1.34	10.42 1.63
Qwen2.5-VL[11]	72B	3.31 2.32	3.08 1.33	13.59 4.50	10.11 2.50	28.08 6.34	27.54 6.10	13.41 4.10	12.17 3.04
MiniCPM-V-2_6[49]	8B	1.25 0.00	1.33 0.00	14.18 2.25	12.64 3.25	29.46 13.90	29.97 12.93	13.09 4.67	12.84 4.67
MiniCPM-O-2_6[50]	8B	1.55 0.00	1.64 0.00	12.26 1.50	13.34 1.50	25.27 11.46	23.60 10.00	11.46 3.75	11.35 3.33
Mantis-Idefics2[52]	8B	0.08 0.00	0.09 0.00	10.24 0.75	9.01 0.25	11.03 0.49	9.73 0.73	6.13 0.35	5.41 0.28
LLaVA-OneVision[53]	7B	1.32 0.00	1.02 0.00	13.45 1.00	15.28 1.25	21.96 6.34	20.98 3.17	10.74 2.12	10.85 1.27
InternVL2.5[59]	8B	0.14 0.00	0.15 0.00	1.67 0.00	1.06 0.00	5.01 0.00	4.44 0.00	1.99 0.00	1.65 0.00
Migician[28]	7B	10.06 5.31	16.97 8.29	16.87 6.75	12.65 5.25	21.73 6.10	25.94 11.22	15.37 5.94	18.35 8.28
HuatuoGPT-Vision[12]	7B	1.24 0.17	1.41 0.00	6.85 0.00	7.95 0.50	12.89 0.73	14.80 1.95	6.21 0.28	7.15 0.71
MedSeq-Grounder (Ours)	7B	82.56 92.04	86.84 96.68	65.54 70.75	89.23 94.50	80.54 94.39	88.01 100.00	77.15 86.69	87.86 97.03

E.4 Failure case study

We conducted a detailed failure analysis to better understand model behavior. Our initial observations show that the most models are able to correctly follow instructions and output coordinates in the required format.

However, as the visual context becomes more complex, model performance drops significantly. For example, Qwen2.5-VL frequently produces bounding boxes that span nearly the entire image ([0, 0, x_max, y_max]), and InternVL3 often outputs predictions that are spatially misaligned with the target region, as illustrated in Fig. 7.

F Potential negative societal impacts

While the proposed benchmark includes eight tasks spanning medical image sequences, the resulting performance is intended for reference purposes only. High scores achieved by MLLMs on MedSG-Bench do not necessarily indicate clinical readiness or real-world applicability. Any deployment in clinical settings requires thorough validation and oversight from qualified medical professionals to ensure safety and reliability.

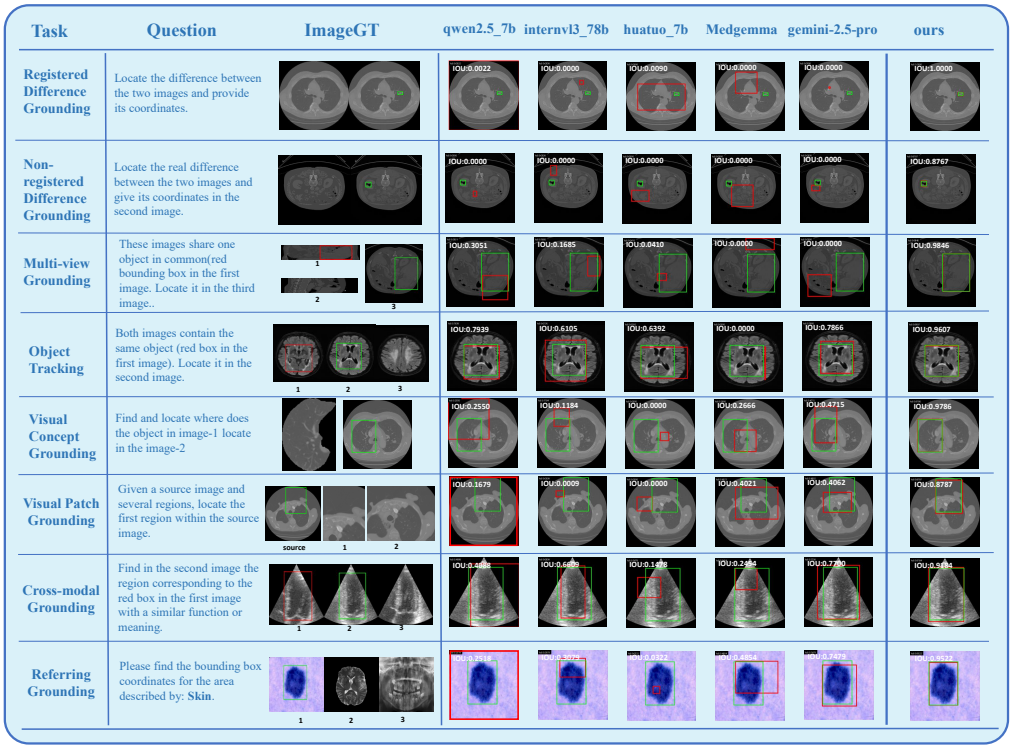

Figure 7: Visualization of samples in MedSG-Bench.