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

NeurIPS 2025 Datasets and Benchmarks Track **spotlight**

Contact: yuejk@bupt.edu.cn
Home Page: https://github.com/Yuejingkun/MedSG-Bench

Table of contents

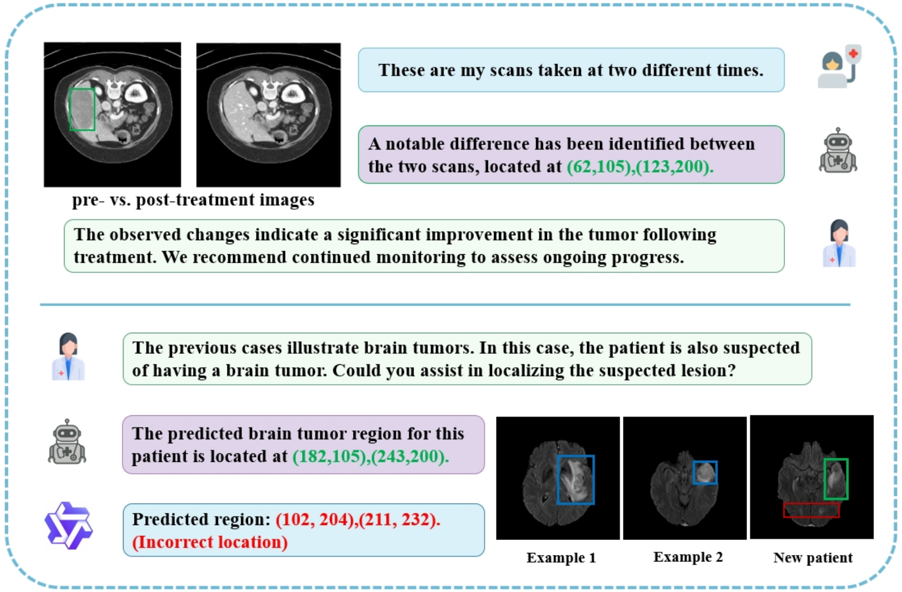
Motivation

- Visual grounding is crucial in medical imaging, where understanding the semantic content of clinical text and accurately localizing the corresponding pathological regions is essential for interpretable and reliable diagnosis.

- Real world clinical diagnosis inherently requires sequential image analysis (e.g., pre- vs. post- treatment images).

- However, existing medical visual grounding benchmarks mainly focus on single-image scenarios.

- To address this gap, we introduce **MedSG-Bench**, the first benchmark tailored for **Med**ical Image **S**equences **G**rounding.



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

Table of contents

MedSG-Bench

- MedSG-Bench is the first benchmark specifically designed to evaluate the grounding capabilities of MLLMs in medical image sequences.

- Construction protocol
  - Data collection
  - Data review and quality filtering
  - Data preprocessing
  - VQA generation

- Data description
  - 76 publicly available datasets
  - 10 medical imaging modalities
  - 114 clinical tasks
  - 9,630 VQA pairs (24,341 medical images)

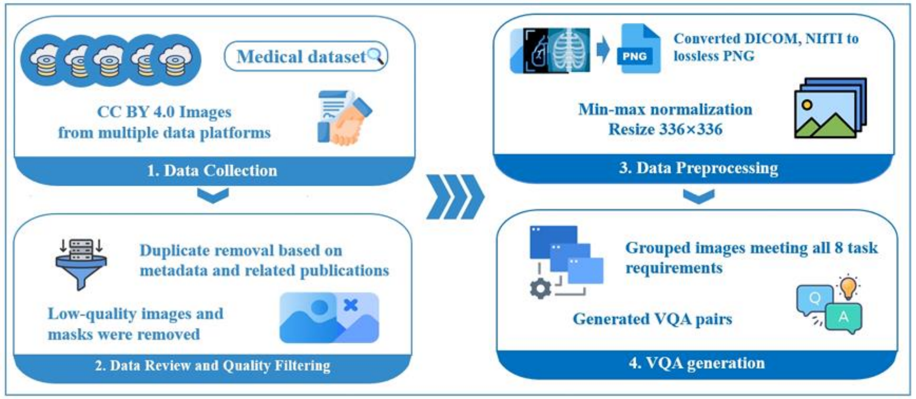


Figure 4: Overview of the MedSG-Bench construction protocol.

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

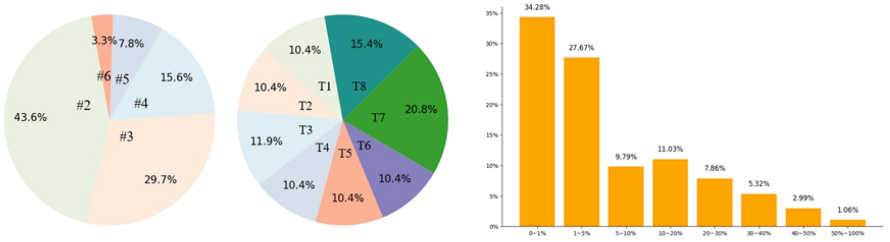
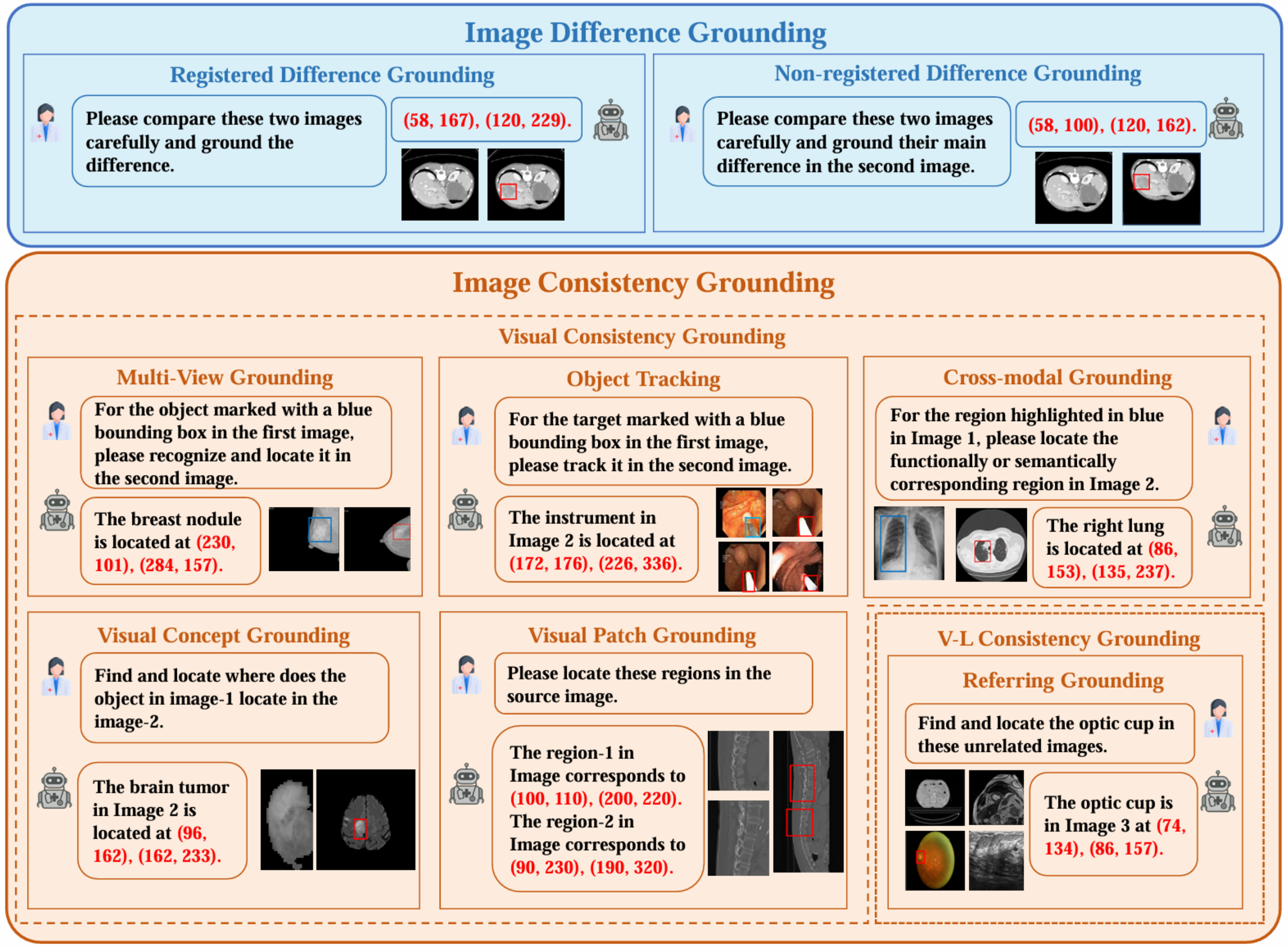


Figure 5: Proportions of image sequence length (**left**), data distribution across tasks (**middle**), and target-to-image size ratios (**right**) in MedSG-Bench.

MedSG-Bench

- MedSG-Bench defines eight tasks grouped into two core paradigms
  - Image **D**ifference **G**rounding (2 tasks)
  - Image **C**onsistency **G**rounding (6 tasks)
    - Visual Consistency Grounding
    - V-L Consistency Grounding



- MedSG-188K & MedSeq-Grounder
  - Ensure diversity and mitigate potential bias by employing multiple LLMs
  - 188,163 VQA samples (324,359 medical images)
  - MedSeq-Grounder is developed based on the Qwen2.5-VL-7B

Table of contents

Evaluation

- We benchmark a diverse collection of MLLMs on MedSG-Bench
  - Proprietary MLLMs
    - GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro
  - General-purpose MLLMs
    - Qwen2.5-VL, InternVL3…
  - Medical-domain specialized MLLMs
    - MedGemma, HuatuoGPT-Vision…
- Zero-shot setting
- Metric: IoU and Acc@0.5

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	<u>28.19</u> <u>23.90</u>	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	<u>14.94</u> <u>10.70</u>	<u>41.11</u> <u>49.20</u>	<u>24.44</u> <u>28.12</u>	<u>28.12</u> <u>22.67</u>	<u>20.66</u> <u>15.61</u>
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	<u>18.69</u> <u>8.59</u>	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	<u>15.26</u> <u>7.80</u>	<u>14.49</u> <u>6.10</u>	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**

Evaluation

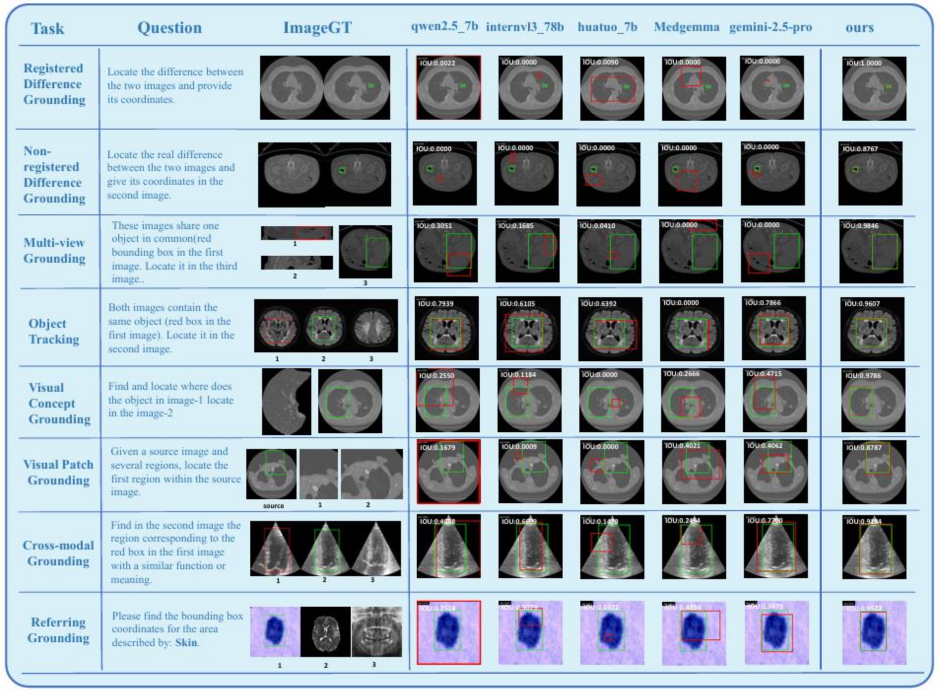
- Finding 1: Grounding in medical image sequences is still challenging for all MLLMs
- Finding 2: All MLLMs exhibit limitations in detecting small medical targets
- Finding 3: Medical-domain specialized models are often worse than general-purpose models
- Finding 4: Larger or newer models do not guarantee improved grounding performance

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**

Evaluation

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

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



- Analysis 1 (left): Effect of clinical windowing on model performance
- Analysis 2 (middle): the potential bias of question generations
- Analysis 3 (right): failure cases visualization

Table of contents

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.

Thanks

NeurIPS 2025 Datasets and Benchmarks Track **spotlight**

Contact: yuejk@bupt.edu.cn

Home Page: https://github.com/Yuejingkun/MedSG-Bench