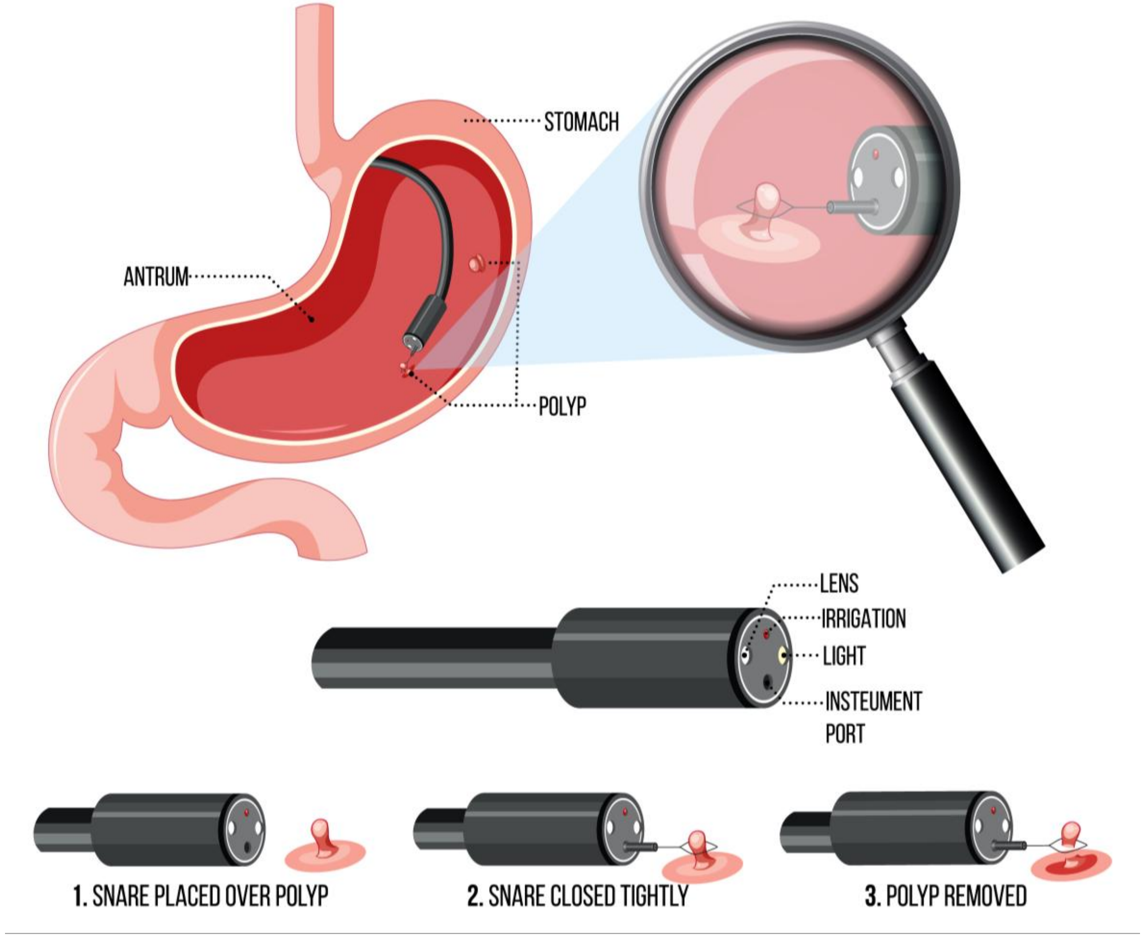
EndoBench: A Comprehensive Evaluation of Multi-Modal Large Language Models for Endoscopy Analysis

1 The Chinese University of Hong Kong
2 The City University of Hong Kong
3 University of Oxford
4 Shanghai AI Laboratory
5 The Sixth Affiliated Hospital, Sun Yat-sen University

Background

Endoscopy's Importance



AI's Impact on Endoscopy

Motivation

Existing Benchmarks in Endoscopy:

- **Lack of multi-scenario evaluation:**
  Typically focus on specific scenarios

- **Lack of multi-dimensional capacities:**
  Ignore the multi-dimensional capacities required in clinical practice

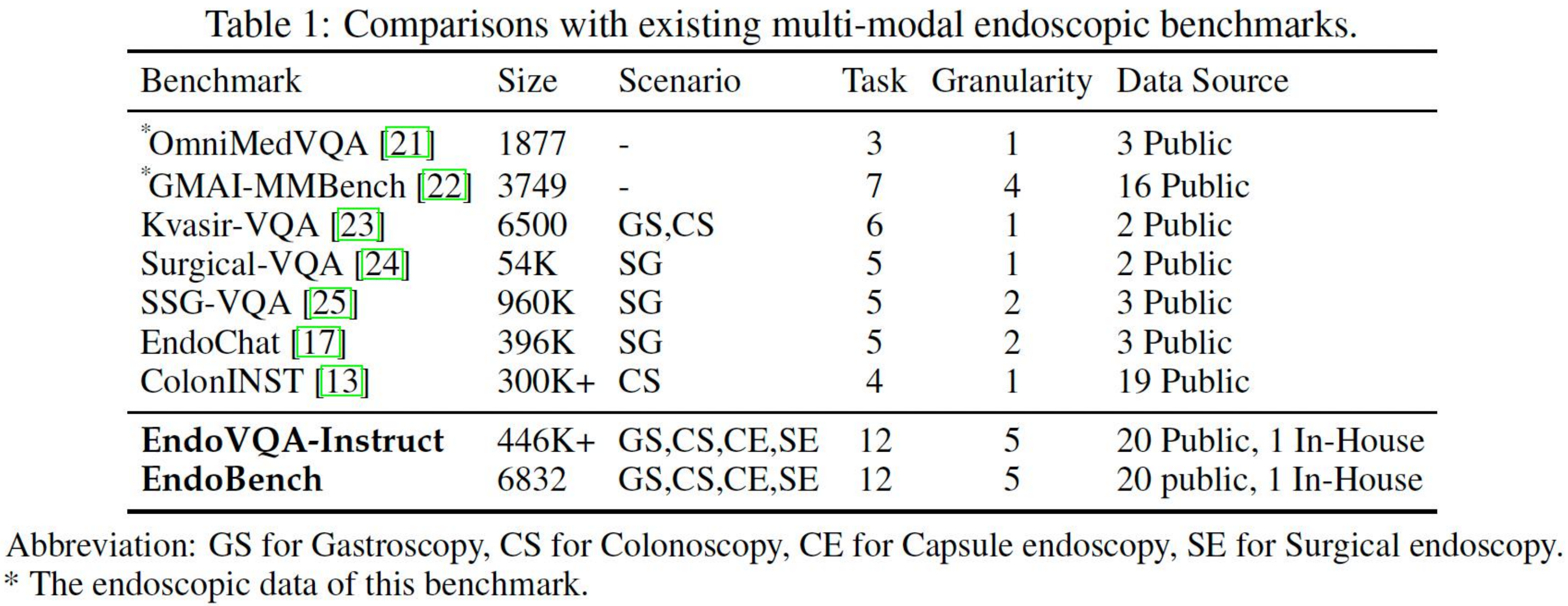


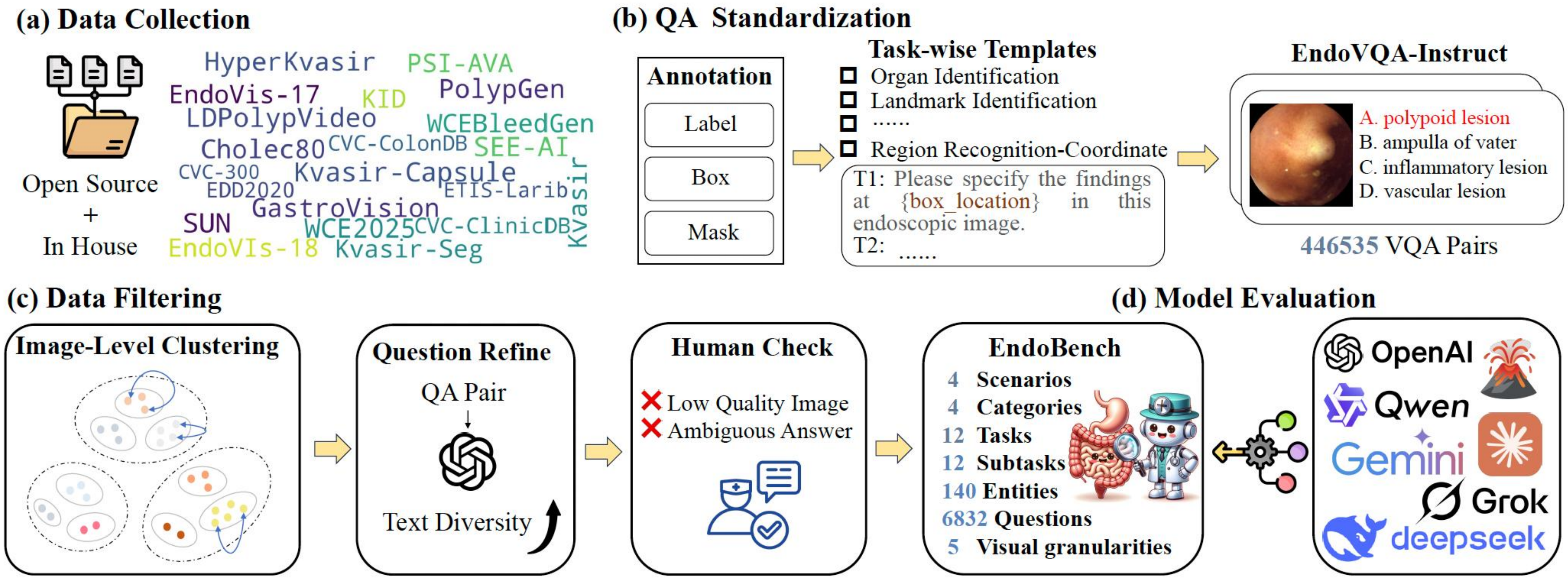
Table 1: Comparisons with existing multi-modal endoscopic benchmarks.

Benchmark	Size	Scenario	Task	Granularity	Data Source
*OmniMedVQA [21]	1877	-	3	1	3 Public
*GMAI-MMBench [22]	3749	-	7	4	16 Public
Kvasir-VQA [23]	6500	GS,CS	6	1	2 Public
Surgical-VQA [24]	54K	SG	5	1	2 Public
SSG-VQA [25]	960K	SG	5	2	3 Public
EndoChat [17]	396K	SG	5	2	3 Public
ColonINST [13]	300K+	CS	4	1	19 Public
EndoVQA-Instruct	446K+	GS,CS,CE,SE	12	5	20 Public, 1 In-House
EndoBench	6832	GS,CS,CE,SE	12	5	20 public, 1 In-House

Abbreviation: GS for Gastroscopy, CS for Colonoscopy, CE for Capsule endoscopy, SE for Surgical endoscopy.
* The endoscopic data of this benchmark.

Data construction

- ➢ Data collection
- ➢ QA Standardization
- ➢ Data Filtering
- ➢ Model Evaluation



(a) Data Collection

Open Source
+
In House

HyperKvasir PSI-AVA EndoVis-17 KID PolypGen LDPolypVideo WCEBleedGen Cholec80 CVC-ColonDB SEE-AI CVC-300 EDD2020 Kvasir-Capsule ETIS-Larib Kvasir GastroVision SUN WCE2025 CVC-ClinicDB EndoVIs-18 Kvasir-Seg

(b) QA Standardization

Annotation

Label

Box

Mask

Task-wise Templates

☐ Organ Identification
☐ Landmark Identification
☐ ……
☐ Region Recognition-Coordinate

T1: Please specify the findings at {box_location} in this endoscopic image.
T2: ……

EndoVQA-Instruct

A. polypoid lesion
B. ampulla of vater
C. inflammatory lesion
D. vascular lesion

446535 VQA Pairs

(c) Data Filtering

Image-Level Clustering

Question Refine

QA Pair

Text Diversity

Human Check

✗ Low Quality Image
✗ Ambiguous Answer

(d) Model Evaluation

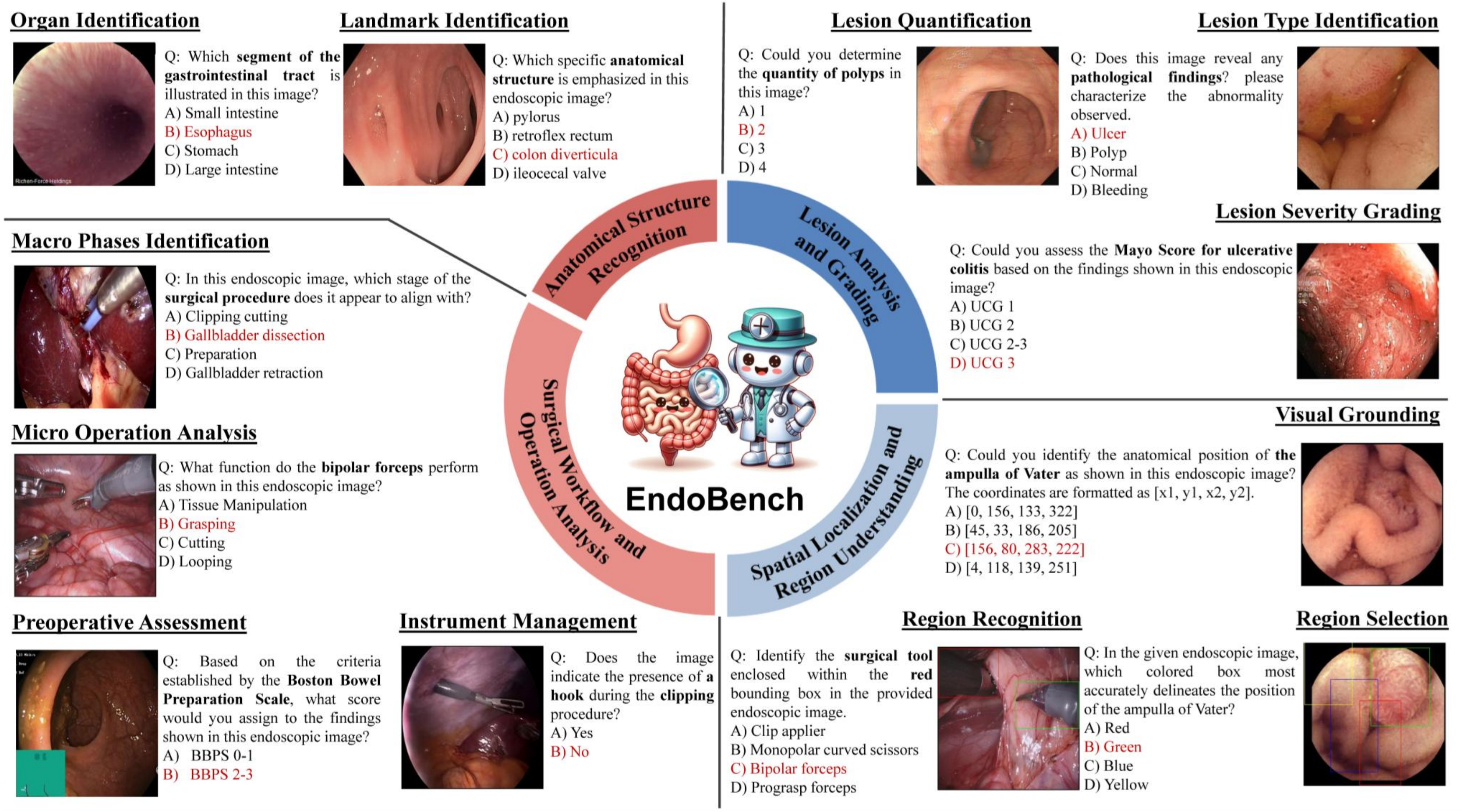
EndoBench

4 Scenarios
4 Categories
12 Tasks
12 Subtasks
140 Entities
6832 Questions
5 Visual granularities

OpenAI
Qwen
Gemini
Grok
deepseek

EndoBench

First Comprehensive Evaluation Benchmark of MLLM in the Endoscopy Domain


Anatomical Structure Recognition
Lesion Analysis and Grading
Surgical Workflow and Operation Analysis
Spatial Localization and Region Understanding
EndoBench
Richen-Force Holdings

Organ Identification

Q: Which **segment of the gastrointestinal tract** is illustrated in this image?
A) Small intestine
B) Esophagus
C) Stomach
D) Large intestine

Landmark Identification

Q: Which specific **anatomical structure** is emphasized in this endoscopic image?
A) pylorus
B) retroflex rectum
C) colon diverticula
D) ileocecal valve

Lesion Quantification

Q: Could you determine the **quantity of polyps** in this image?
A) 1
B) 2
C) 3
D) 4

Lesion Type Identification

Q: Does this image reveal any **pathological findings**? please characterize the abnormality observed.
A) Ulcer
B) Polyp
C) Normal
D) Bleeding

Lesion Severity Grading

Q: Could you assess the **Mayo Score for ulcerative colitis** based on the findings shown in this endoscopic image?
A) UCG 1
B) UCG 2
C) UCG 2-3
D) UCG 3

Macro Phases Identification

Q: In this endoscopic image, which stage of the **surgical procedure** does it appear to align with?
A) Clipping cutting
B) Gallbladder dissection
C) Preparation
D) Gallbladder retraction

Micro Operation Analysis

Q: What function do the **bipolar forceps** perform as shown in this endoscopic image?
A) Tissue Manipulation
B) Grasping
C) Cutting
D) Looping

Visual Grounding

Q: Could you identify the anatomical position of **the ampulla of Vater** as shown in this endoscopic image? The coordinates are formatted as [x1, y1, x2, y2].
A) [0, 156, 133, 322]
B) [45, 33, 186, 205]
C) [156, 80, 283, 222]
D) [4, 118, 139, 251]

Preoperative Assessment

Q: Based on the criteria established by the **Boston Bowel Preparation Scale**, what score would you assign to the findings shown in this endoscopic image?
A) BBPS 0-1
B) BBPS 2-3

Instrument Management

Q: Does the image indicate the presence of **a hook** during the **clipping** procedure?
A) Yes
B) No

Region Recognition

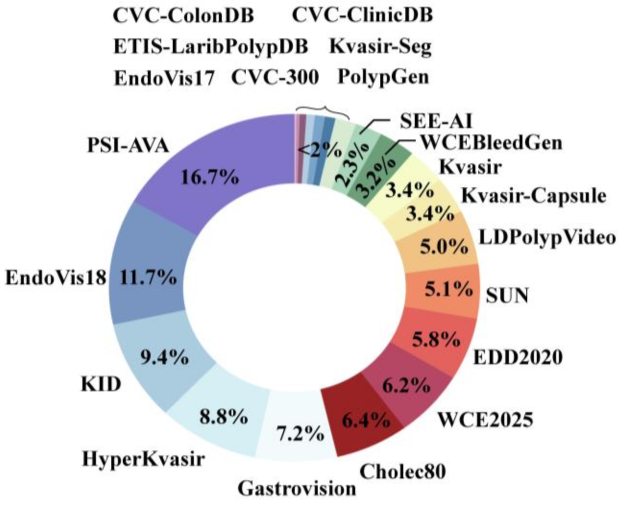
Q: Identify the **surgical tool** enclosed within the **red** bounding box in the provided endoscopic image.
A) Clip applier
B) Monopolar curved scissors
C) Bipolar forceps
D) Prograsp forceps

Region Selection

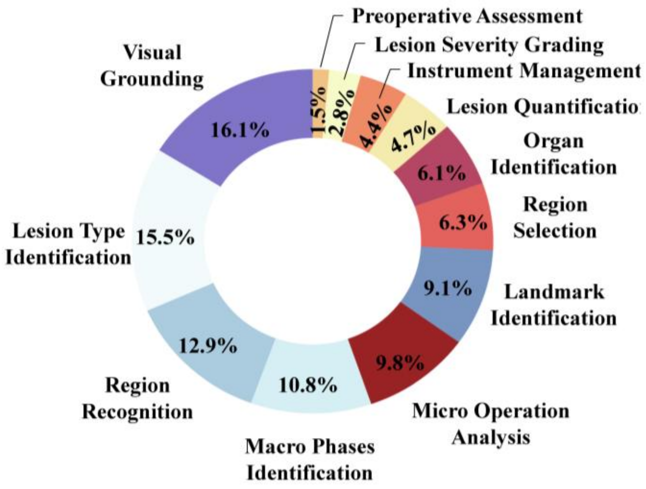
Q: In the given endoscopic image, which colored box most accurately delineates the position of the ampulla of Vater?
A) Red
B) Green
C) Blue
D) Yellow

Data Distribution

- ➢ 21 Datasets
- ➢ 4 Scenarios
- ➢ 12 Tasks with 12 Subtasks
- ➢ 140 Entities
- ➢ 5 Visual granularities



(b) Dataset Distribution



(c) Task Distribution

Table 4: Statistics regarding the endoscopic scenarios and dataset information covered by the datasets involved.

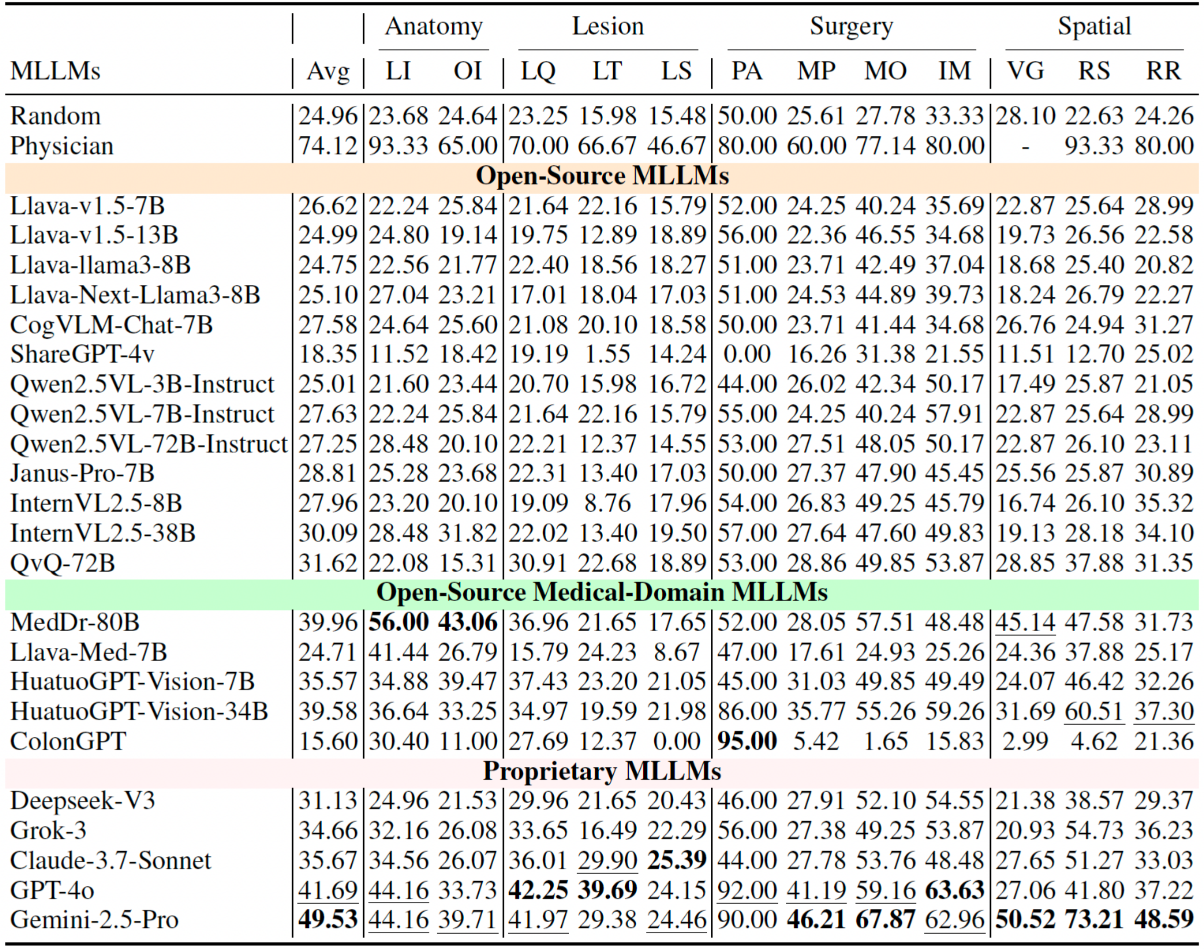
Index	Name	Scenario	Num	Task	Access
1	Kvasir	GS, CS, SE	8000	Classification	Open Access
2	HyperKvasir	GS, CS, SE	10662	Classification	Open Access
3	Kvasir-Capsule	CE	47238	Classification	Open Access
4	GastroVision	GS, CS, SE	8000	Classification	Open Access
5	KID	CE	2371	Classification, Segmentation	Restricted Access
6	WCEBleedGen	CE	2618	Classification, Segmentation	Open Access
7	SEE-AI	GS	18481	Classification	Open Access
8	Kvasir-Seg	CS	1000	Segmentation	Open Access
9	CVC-ColonDB	CS	380	Segmentation	Open Access
10	ETIS-Larib	CS	196	Segmentation	Open Access
11	CVC-ClinicDB	CS	612	Segmentation	Open Access
12	CVC-300	CS	60	Segmentation	Open Access
13	EDD2020	CS	386	Classification, Segmentation	Open Access
14	SUN-Database	CS	130519	Classification, Segmentation	Restricted Access
15	LDPolypVideo	CS	40266	Detection	Open Access
16	PolypGen	CS	8037	Segmentation	Open Access
17	Cholec80	SE	8080	Classification, Segmentation	Open Access
18	EndoVis-17	SE	2235	Classification, Segmentation	Open Access
19	EndoVis-18	SE	2400	Classification, Segmentation	Open Access
20	PSI-AVA	SE	4471	Classification, Segmentation	Open Access
21	WCE2025	CE	23447	Classification	In House

Abbreviation: GS for Gastroscopy, CS for Colonoscopy, CE for Capsule endoscopy, SE for Surgical endoscopy.

Model Evaluation

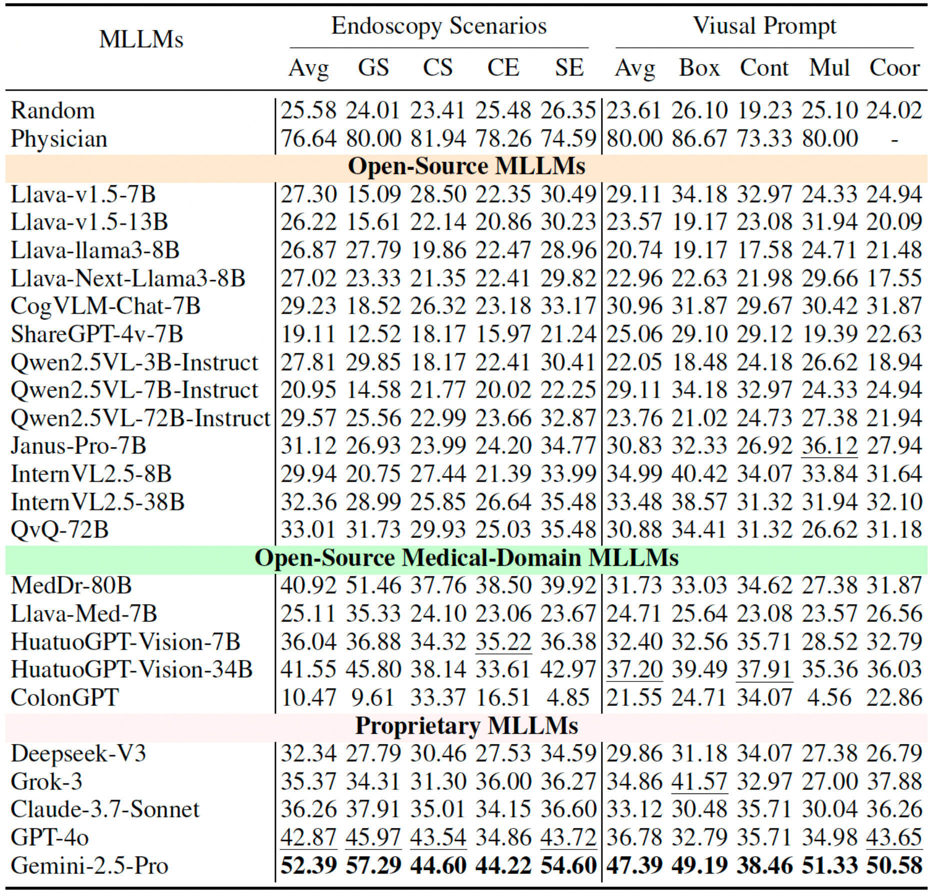
- Benchmarking 23 state-of-the-art models, including general-purpose, medical-specialized, and proprietary MLLMs

Table 2: Results of different MLLMs on 12 clinical tasks in EndoBench. The best-performing model in each category is **in-bold**, and the second best is underlined.



		Anatomy		Lesion			Surgery				Spatial		
MLLMs	Avg	LI	OI	LQ	LT	LS	PA	MP	MO	IM	VG	RS	RR
Random	24.96	23.68	24.64	23.25	15.98	15.48	50.00	25.61	27.78	33.33	28.10	22.63	24.26
Physician	74.12	93.33	65.00	70.00	66.67	46.67	80.00	60.00	77.14	80.00	-	93.33	80.00
Open-Source MLLMs													
Llava-v1.5-7B	26.62	22.24	25.84	21.64	22.16	15.79	52.00	24.25	40.24	35.69	22.87	25.64	28.99
Llava-v1.5-13B	24.99	24.80	19.14	19.75	12.89	18.89	56.00	22.36	46.55	34.68	19.73	26.56	22.58
Llava-llama3-8B	24.75	22.56	21.77	22.40	18.56	18.27	51.00	23.71	42.49	37.04	18.68	25.40	20.82
Llava-Next-Llama3-8B	25.10	27.04	23.21	17.01	18.04	17.03	51.00	24.53	44.89	39.73	18.24	26.79	22.27
CogVLM-Chat-7B	27.58	24.64	25.60	21.08	20.10	18.58	50.00	23.71	41.44	34.68	26.76	24.94	31.27
ShareGPT-4v	18.35	11.52	18.42	19.19	1.55	14.24	0.00	16.26	31.38	21.55	11.51	12.70	25.02
Qwen2.5VL-3B-Instruct	25.01	21.60	23.44	20.70	15.98	16.72	44.00	26.02	42.34	50.17	17.49	25.87	21.05
Qwen2.5VL-7B-Instruct	27.63	22.24	25.84	21.64	22.16	15.79	55.00	24.25	40.24	57.91	22.87	25.64	28.99
Qwen2.5VL-72B-Instruct	27.25	28.48	20.10	22.21	12.37	14.55	53.00	27.51	48.05	50.17	22.87	26.10	23.11
Janus-Pro-7B	28.81	25.28	23.68	22.31	13.40	17.03	50.00	27.37	47.90	45.45	25.56	25.87	30.89
InternVL2.5-8B	27.96	23.20	20.10	19.09	8.76	17.96	54.00	26.83	49.25	45.79	16.74	26.10	35.32
InternVL2.5-38B	30.09	28.48	31.82	22.02	13.40	19.50	57.00	27.64	47.60	49.83	19.13	28.18	34.10
QvQ-72B	31.62	22.08	15.31	30.91	22.68	18.89	53.00	28.86	49.85	53.87	28.85	37.88	31.35
Open-Source Medical-Domain MLLMs													
MedDr-80B	39.96	**56.00**	**43.06**	36.96	21.65	17.65	52.00	28.05	57.51	48.48	45.14	47.58	31.73
Llava-Med-7B	24.71	41.44	26.79	15.79	24.23	8.67	47.00	17.61	24.93	25.26	24.36	37.88	25.17
HuatuoGPT-Vision-7B	35.57	34.88	39.47	37.43	23.20	21.05	45.00	31.03	49.85	49.49	24.07	46.42	32.26
HuatuoGPT-Vision-34B	39.58	36.64	33.25	34.97	19.59	21.98	86.00	35.77	55.26	59.26	31.69	60.51	37.30
ColonGPT	15.60	30.40	11.00	27.69	12.37	0.00	**95.00**	5.42	1.65	15.83	2.99	4.62	21.36
Proprietary MLLMs													
Deepseek-V3	31.13	24.96	21.53	29.96	21.65	20.43	46.00	27.91	52.10	54.55	21.38	38.57	29.37
Grok-3	34.66	32.16	26.08	33.65	16.49	22.29	56.00	27.38	49.25	53.87	20.93	54.73	36.23
Claude-3.7-Sonnet	35.67	34.56	26.07	36.01	29.90	**25.39**	44.00	27.78	53.76	48.48	27.65	51.27	33.03
GPT-4o	41.69	44.16	33.73	**42.25**	**39.69**	24.15	92.00	41.19	59.16	**63.63**	27.06	41.80	37.22
Gemini-2.5-Pro	**49.53**	44.16	39.71	41.97	29.38	24.46	90.00	**46.21**	**67.87**	62.96	**50.52**	**73.21**	**48.59**

Table 3: Results of different MLLMs on 4 different endoscopy scenarios and 4 different visual prompts in EndoBench. The best-performing model in each category is **in-bold**, and the second best is underlined.



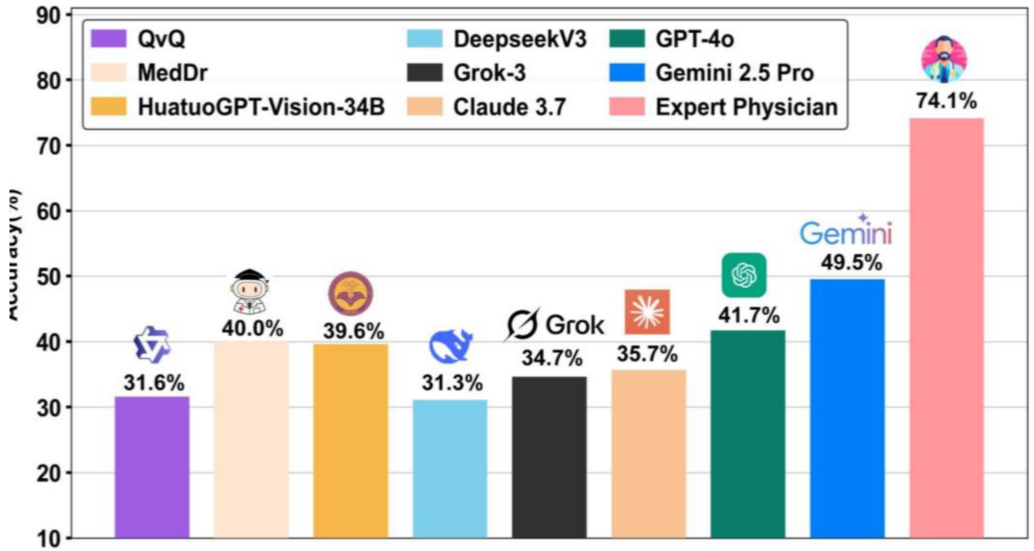
MLLMs	Endoscopy Scenarios					Viusal Prompt				
	Avg	GS	CS	CE	SE	Avg	Box	Cont	Mul	Coor
Random	25.58	24.01	23.41	25.48	26.35	23.61	26.10	19.23	25.10	24.02
Physician	76.64	80.00	81.94	78.26	74.59	80.00	86.67	73.33	80.00	-
Open-Source MLLMs										
Llava-v1.5-7B	27.30	15.09	28.50	22.35	30.49	29.11	34.18	32.97	24.33	24.94
Llava-v1.5-13B	26.22	15.61	22.14	20.86	30.23	23.57	19.17	23.08	31.94	20.09
Llava-llama3-8B	26.87	27.79	19.86	22.47	28.96	20.74	19.17	17.58	24.71	21.48
Llava-Next-Llama3-8B	27.02	23.33	21.35	22.41	29.82	22.96	22.63	21.98	29.66	17.55
CogVLM-Chat-7B	29.23	18.52	26.32	23.18	33.17	30.96	31.87	29.67	30.42	31.87
ShareGPT-4v-7B	19.11	12.52	18.17	15.97	21.24	25.06	29.10	29.12	19.39	22.63
Qwen2.5VL-3B-Instruct	27.81	29.85	18.17	22.41	30.41	22.05	18.48	24.18	26.62	18.94
Qwen2.5VL-7B-Instruct	20.95	14.58	21.77	20.02	22.25	29.11	34.18	32.97	24.33	24.94
Qwen2.5VL-72B-Instruct	29.57	25.56	22.99	23.66	32.87	23.76	21.02	24.73	27.38	21.94
Janus-Pro-7B	31.12	26.93	23.99	24.20	34.77	30.83	32.33	26.92	36.12	27.94
InternVL2.5-8B	29.94	20.75	27.44	21.39	33.99	34.99	40.42	34.07	33.84	31.64
InternVL2.5-38B	32.36	28.99	25.85	26.64	35.48	33.48	38.57	31.32	31.94	32.10
QvQ-72B	33.01	31.73	29.93	25.03	35.48	30.88	34.41	31.32	26.62	31.18
Open-Source Medical-Domain MLLMs										
MedDr-80B	40.92	51.46	37.76	38.50	39.92	31.73	33.03	34.62	27.38	31.87
Llava-Med-7B	25.11	35.33	24.10	23.06	23.67	24.71	25.64	23.08	23.57	26.56
HuatuoGPT-Vision-7B	36.04	36.88	34.32	35.22	36.38	32.40	32.56	35.71	28.52	32.79
HuatuoGPT-Vision-34B	41.55	45.80	38.14	33.61	42.97	37.20	39.49	37.91	35.36	36.03
ColonGPT	10.47	9.61	33.37	16.51	4.85	21.55	24.71	34.07	4.56	22.86
Proprietary MLLMs										
Deepseek-V3	32.34	27.79	30.46	27.53	34.59	29.86	31.18	34.07	27.38	26.79
Grok-3	35.37	34.31	31.30	36.00	36.27	34.86	41.57	32.97	27.00	37.88
Claude-3.7-Sonnet	36.26	37.91	35.01	34.15	36.60	33.12	30.48	35.71	30.04	36.26
GPT-4o	42.87	45.97	43.54	34.86	43.72	36.78	32.79	35.71	34.98	43.65
Gemini-2.5-Pro	**52.39**	**57.29**	**44.60**	**44.22**	**54.60**	**47.39**	**49.19**	**38.46**	**51.33**	**50.58**

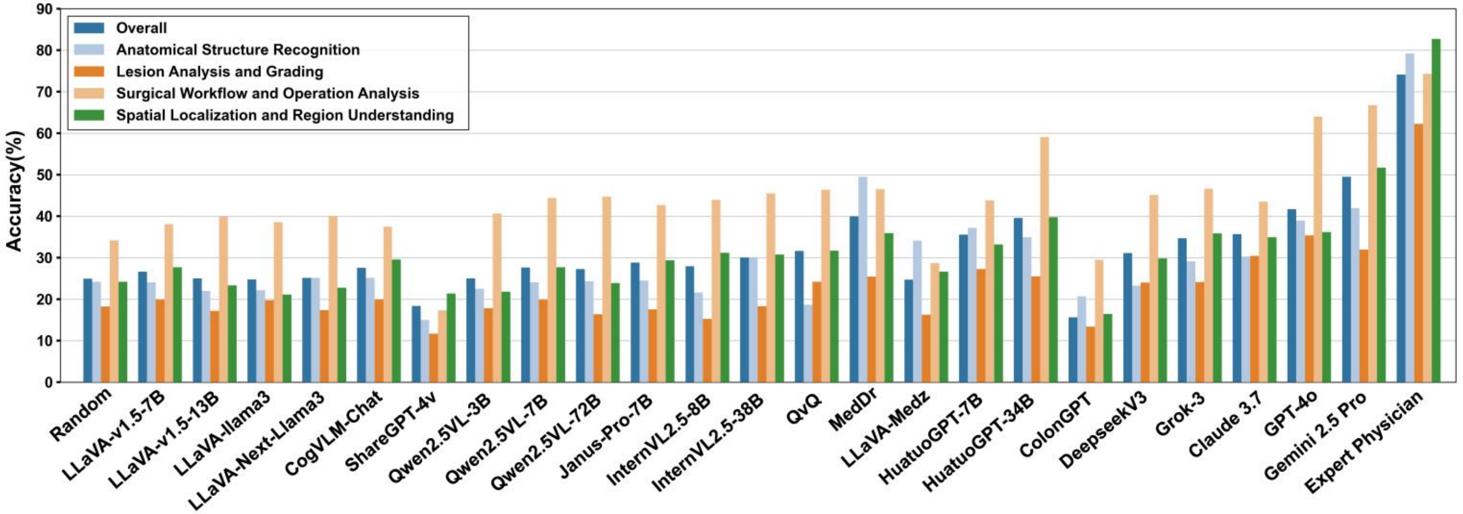
Abbreviation: Cont for Contour, Mul for Multi-region, Coor for Coordinate.

Observation

Observation 1: Endoscopy remains a challenging domain for MLLMs, with significant gaps between models and human expertise

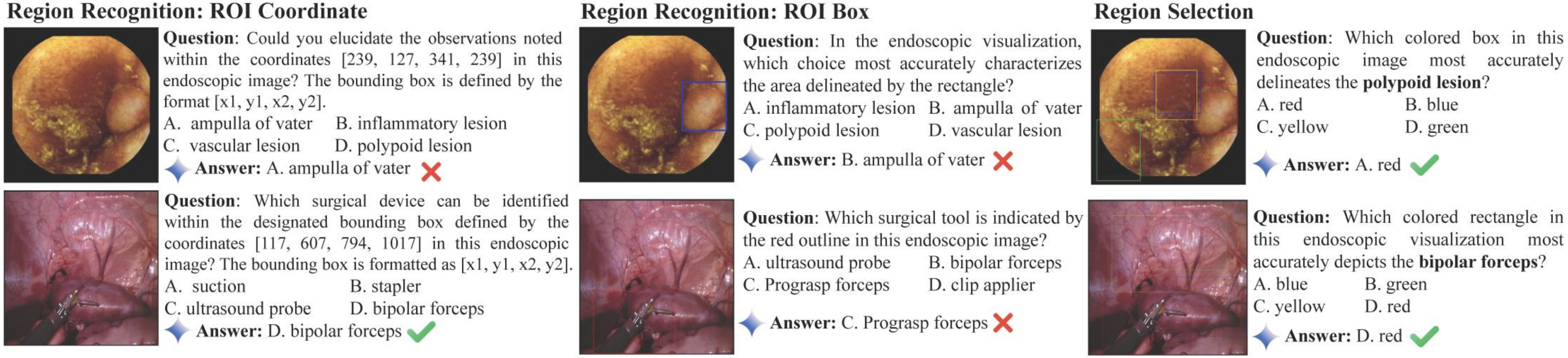
Observation 2: Medical domain-specific Supervised Fine-Tuning markedly boosts model performance





Observation

Observation 3: Model performance varies with visual prompt formats, exposing a gap between visual perception and medical comprehension



Region Recognition: ROI Coordinate

Question: Could you elucidate the observations noted within the coordinates [239, 127, 341, 239] in this endoscopic image? The bounding box is defined by the format [x1, y1, x2, y2].
A. ampulla of vater B. inflammatory lesion
C. vascular lesion D. polypoid lesion
Answer: A. ampulla of vater ✗

Question: Which surgical device can be identified within the designated bounding box defined by the coordinates [117, 607, 794, 1017] in this endoscopic image? The bounding box is formatted as [x1, y1, x2, y2].
A. suction B. stapler
C. ultrasound probe D. bipolar forceps
Answer: D. bipolar forceps ✓

Region Recognition: ROI Box

Question: In the endoscopic visualization, which choice most accurately characterizes the area delineated by the rectangle?
A. inflammatory lesion B. ampulla of vater
C. polypoid lesion D. vascular lesion
Answer: B. ampulla of vater ✗

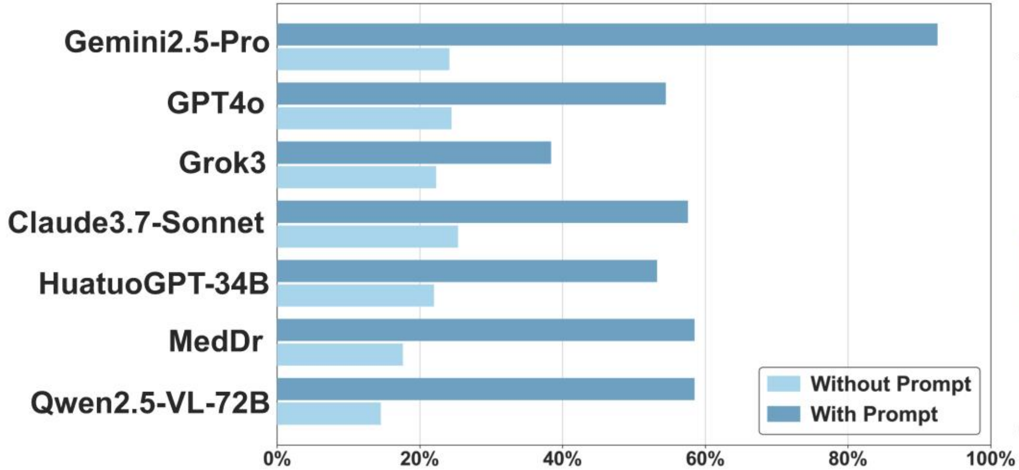
Question: Which surgical tool is indicated by the red outline in this endoscopic image?
A. ultrasound probe B. bipolar forceps
C. Prograsp forceps D. clip applier
Answer: C. Prograsp forceps ✗

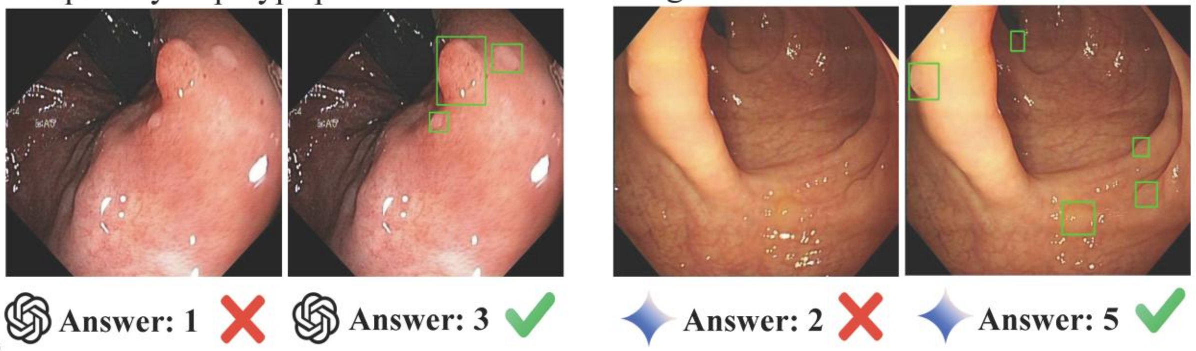
Region Selection

Question: Which colored box in this endoscopic image most accurately delineates the **polypoid lesion**?
A. red B. blue
C. yellow D. green
Answer: A. red ✓

Question: Which colored rectangle in this endoscopic visualization most accurately depicts the **bipolar forceps**?
A. blue B. green
C. yellow D. red
Answer: D. red ✓

Observation 4: Polyp counting exposes dual challenges in lesion identification and numerical reasoning





Question: Based on the endoscopic visualization provided, could you assess the quantity of polyps present?

Answer: 1 ✗ Answer: 3 ✓

Question: How many polyps can be discerned in the provided endoscopic image?

Answer: 2 ✗ Answer: 5 ✓

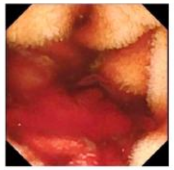
Error Analysis

- Perceptual Errors
- Lack of Knowledge
- Irrelevant Response
- Refusal to Answer

Lack of Knowledge

Question: Does this endoscopic image reveal any pathological findings? If so, could you identify the specific type?

Options:
A. Angiectasia
B. Polyp
C. Bleeding
D. Erosion



HuatuoGPT-Vision-34B Response:

To analyze the endoscopic image and identify any pathological findings, let's consider each option:

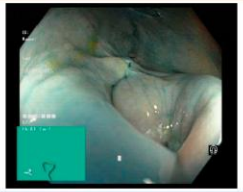
A. Angiectasia: Angiectasias are small dilated blood vessels that can appear as red spots or patches on mucosal surfaces in certain conditions. They often have a characteristic appearance of tiny red dots or blotches.
B. Polyp: Polyps are growths protruding from the lining of organs such as the colon, stomach, or nasal passages.
C. Bleeding: Bleeding would be indicated by fresh blood visible within the lumen (interior space) of the organ being examined.
D. Erosion: Erosions refer to superficial breaks or ulcerations in the mucosa which might result from various causes including inflammation, infection, or physical trauma.

Upon examining the provided image: There is no clear evidence of angiectasia due to their typical presentation. No well-defined mass consistent with a polyp is observed. **While there appears to be some reddish discoloration, it does not clearly indicate active bleeding.** However, the disrupted and irregular tissue texture suggests possible erosion. Based on these observations, the most likely pathological finding depicted in this endoscopic image is: D. erosion

Perceptual Errors

Question: Could you determine the therapeutic procedure depicted in this image?

Options:
A. Dyed lifted polyps
B. Resection margins
C. Resected polyps
D. Dyed resection margins



HuatuoGPT-Vision-34B Response:

To determine the therapeutic procedure depicted in this image, let's analyze each option step by step:

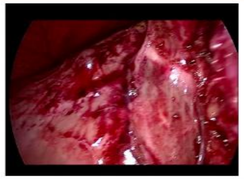
A. dyed lifted polyps - This suggests that there are polyps present which have been stained (dyed) and then elevated for removal or examination. The presence of a dye might be indicated if we see distinct coloration on certain areas within the tissue. **However, no such staining is evident here.**
B. resection margins - Resection margins refer to the edges of the resected area where cancerous cells may still remain after surgery. They would typically not appear as clearly defined structures unless marked with surgical tools or dyes.
C. resected polyps - If polyps were removed from the mucosal surface, one could expect to see clear cuts or disruptions along their former locations. There doesn’t seem to be any obvious signs of recent resection marks like bleeding spots or exposed submucosa layers.
D. dyed resection indicates that the margins are highlighted using dye. Again, without visible evidence of specific dye application, this seems less likely.

Given the absence of distinctive features associated with options A, C, and D, the most plausible answer based on what can be observed in the provided image appears to align more closely with option B, focusing on examining the potential remaining margins post-resection rather than highlighting them through dye or showing actual resections.

Irrelevant Response

Question: Based on the endoscopic visualization, could you ascertain the number of surgical instruments depicted?

Options:
A. 0
B. 1
C. 2
D. 3



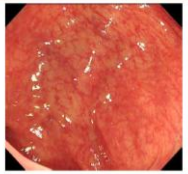
LLaVA-Med Response:

The image shows the number of surgical instruments depicted in the endoscopic visualization. **To determine the exact number, please select the correct answer from the options provided.**

Refusal to Answer

Question: Could you identify the coordinates of the low-grade adenoma depicted in this endoscopic image? The required format for the response is [x1, y1, x2, y2].

Options:
A. [251, 480, 343, 586]
B. [529, 392, 617, 488]
C. [488, 838, 584, 954]
D. [884, 149, 977, 250]



GPT-4o Response:

I'm sorry, **but I'm unable to identify the coordinates of the low-grade adenoma** from the image provided.

The End

Thanks for your time!

EndoBench: A Comprehensive Evaluation of Multi-Modal Large Language Models for Endoscopy Analysis

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu and Yixuan Yuan

Project Page

Group Web