MedHorizon: Towards Long-context Medical Video Understanding in the Wild

Bodong Du[1]† Bowen Liu[1]† Yang Yu[1] Xinpeng Ding[3] Zhiheng Wu[2] Shuning Wang[2]
Shuo Nie[2] Naiming Liu[2] Qifeng Chen[1] Yangqiu Song[1] Xiaomeng Li[1]*

† Equal contribution * Corresponding author

[1] The Hong Kong University of Science and Technology [2] Baidu Inc. [3] Xidian University

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
Baidu 百度
西安电子科技大学 XIDIAN UNIVERSITY
NEURAL INFORMATION PROCESSING SYSTEMS

Retrieve sparse evidence. Reason across full procedures.

An in-the-wild benchmark for medical video MLLMs

FULL-LENGTH VIDEOS	DECODED PROCEDURES	EVIDENCE-GROUNDED QAs	MEAN EVIDENCE RATIO
340	759 h	1,253	0.166%
8 public video sources	7 organs · surgery + examination	13 clinical tasks	≈ 1.7 evidence frames per 1,000

01 Why MedHorizon?

Clinical review spans hours. Decisive findings are sparse, spatially subtle, and dependent on procedural context.

Retrieval comes before reasoning.

Evidence-grounded construction

1 Curate complete procedures

8 public sources cover surgery and examination, including colonoscopy, capsule endoscopy, and ultrasound.

2 Generate from explicit evidence

Use expert annotations; filter ambiguous items, balance answers, rewrite questions, and strengthen distractors.

3 Verify with human review

Audit video usability, QA quality, evidence support, and consistency with the original annotation.

13 tasks · two capability levels

Fine-grained understanding (8)

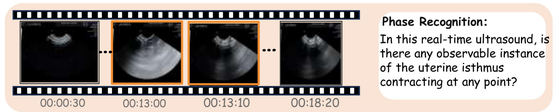
Phase, step, region, bowel quality; largest lesion, site, histology, and size.

Semantic reasoning (5)

From instrument recognition to multi-hop aggregation.

02 Clinical task demonstrations

Fine-grained Understanding



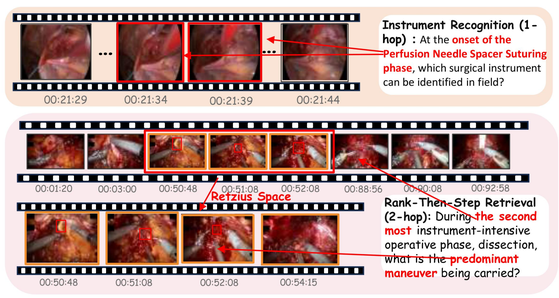
Phase Recognition
Is uterine isthmus contraction visible?

Region Localization
Which GI region is shown at 3 h 01–05 min?

Step Discrimination
Which surgical step occurs at 2952–2982 s?

Histology & Size Binding
What are the first lesion's histology and size?

Semantic Reasoning



1-hop · Recognize

Identify the instrument at a specified phase onset.

2-hop · Compare + retrieve

Rank phases by instrument use, then retrieve the maneuver.

3-hop · Bind + aggregate

Find lesions, bind them to colonic segments, then count.

03 Results & diagnostics

Overall accuracy on the full benchmark

Model	Family	Acc. %
Gemini-3.1-Pro-Preview	**Closed**	**41.1**
GPT-5.4	Closed	**32.6**
LLaVA-Video-72B	General	**29.4**
WFS-SB	Long video	**28.1**
Lingshu-32B	Medical	**27.9**
Qwen3-VL-8B	General	**26.2**

Selected baselines · original model-specific input budgets

Human experts: 96.9% on a preliminary subset
Not directly comparable with full-benchmark model scores.

What limits current models?

RQ2 · Oracle evidence helps, but is insufficient

Semantic accuracy: uniform → verified GT-images

MedGRPO	13.6% →	33.9%
Lingshu-7B	22.9% →	38.0%

RQ1 More frames do not reliably help.
Extra context may dilute sparse evidence.

RQ3 Weak chronology + attention drift.
Qwen3-VL-8B: 26.5 → 24.2–24.5% after order disruption.

RQ4 Local detail vs. global coverage.
Generic sampling yields modest, task-dependent gains.

Code & data: github.com/xmed-lab/MedHorizon