LLaVA-Endo: A Large Language-and-Vision Assistant for Gastrointestinal Endoscopy

Jieru YAO, Xueran LI, Qiang XIE, Longfei HAN, Yiwen JIA,

Nian LIU, Dingwen ZHANG, Junwei HAN

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40319-8

Problems & Ideas

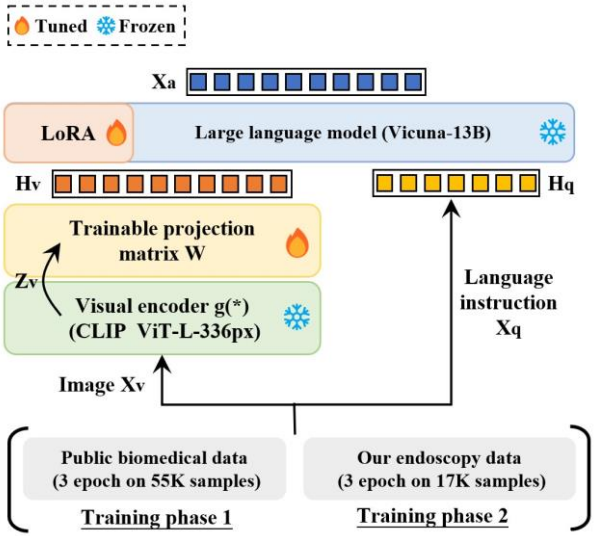
- **Problems**

  - During GI endoscopic examinations, physicians may encounter practical hindrances, fatigue, stress, or limited experience, which can lead to erroneous results.

  - Existing AI-assisted GI endoscopy methods rely on different visual models to perform various tasks independently, leading to restricted interactivity and adaptability to user instructions.

- **Ideas**

  - We propose Large Language and Vision Assistant for Gastrointestinal Endoscopy (LLaVA-Endo), a first attempt to extend multi-modal instruction-tuning to GI endoscopy.



The LLaVA-Endo framework, which is a large language and vision assistant for the GI endoscopy diagnosis.

Main Contributions

- **Contributions:**

-We introduce LLaVA-Endo, which is the first large language and vision model fine-tuned with GI endoscopic knowledge.

- We generate a large-scale multi-modal instruction-following GI endoscopic datasets using examination reports.

- We propose a novel progressive transfer learning method to ensure accurate and domain-specific knowledge in LLaVA-Endo.

- Experiments on held-out evaluation dataset outperform state-of-the-art (SoTA) baselines and can effectively handle a diverse set of tasks.

GI endoscopic visual instruction-following example

Caption:

The image shows a polypoid protrusion in the rectum, measuring approximately 0.5cm × 0.5cm, with a pedicle and a smooth surface.

Conversation:

User: Does the input image have any abnormalities?

Assistant: The image shows the rectum with a polypoid protrusion measuring approximately 0.5cm × 0.5cm, with a pedicle, and a smooth surface.

User: What causes the formation of the polyp?

Assistant: The formation of polyps is attributed to a combination of various factors, including genetic factors, prolonged chronic inflammation, unhealthy lifestyle habits, lack of physical activity and so on.

Example of LLaVA-Endo medical visual chat and reasoning capabilities. The original caption is also given for reference.

Methods	Affiliation	Scoring ↑				
		GPT-4	Human 1	Human 2	Human 3	Average
GPT-4V	OpenAI	<u>8.56</u>	<u>5.85</u>	<u>5.16</u>	5.02	<u>6.15</u>
Gemini	Google	2.59	3.73	3.86	4.05	3.56
LLaVA-med-7b	Microsoft	3.99	4.02	3.32	4.05	3.85
LLaVA1.5-7b	Microsoft	5.06	4.32	4.43	5.30	4.78
LLaVA1.5-13b	Microsoft	5.16	5.28	4.99	<u>5.68</u>	5.28
mPLUG-Owl	DAMO	4.67	2.54	2.60	2.59	3.10
MiniGPT-v2	Vision-CAIR	3.44	2.35	2.04	2.36	2.55
LLaVA-Endo	WisOmni-Tech	**9.19**	**7.93**	**9.30**	**7.85**	**8.57**

Comparison of our method and SoTA methods. Best results in bold, second best in underline.