Foundational Multi-Task Multimodal Model for Upper GI Endoscopy*

Yuxuan He
University of Massachusetts Lowell
Yuxuan_He@student.uml.edu

Qilei Chen
University of Massachusetts Lowell
Qilei_Chen1@uml.edu

Benyuan Liu
University of Massachusetts Lowell
benyuan_liu@uml.edu

Yu Cao
University of Massachusetts Lowell
Yu_Cao@uml.edu

Abstract

Upper gastrointestinal (GI) diseases are a major health threat. Especially gastric cancer, which is the 5th most common cancer worldwide and the 3rd common cause of cancer death. In the U.S., around 62 million people are diagnosed with digestive disorders annually. Gastroscopy is currently the only commonly recognized effective method for screening and treatment. Even though there are many machine learning-based techniques for computer-aided detection and diagnosis, multimodal models that integrate both textual and visual information are extremely lacking. In this paper, we present a foundational multimodal framework for upper GI endoscopy, adapting recent vision-language model architectures. For the first time, we have applied large-scale multimodal modeling in the upper gastrointestinal (GI) endoscopy field, which is a domain that has received less attention in prior medical image research. Based on the four downstream tasks established in previous work, our framework integrates both public and internal GI endoscopy expert-annotated datasets. Through the unified instruction-based fine-tuning pipeline, the framework supports most upper GI diagnostic, detection, and reporting scenarios. It not only covers clinical tasks such as anatomical recognition, lesion diagnosis, and lesion localization, but also enables natural language interaction and communication with clinicians, greatly enhancing the model's practicality and clinical value. Experimental results show that the framework can generalization and perform in various GI endoscopy tasks. It is able to handle a variety of medical tasks and report summarization capabilities that meet the needs of medical experts. Although it still has space for improvement in some tasks, the framework lays the foundation for further optimization and wider application.

1. Introduction

Upper gastrointestinal (Upper GI) diseases are a major health threat in worldwide, like gastric cancer as an example, which is the fifth most common cancer in the world and the third leading cause of cancer death [1]. In the United States, about 62 million people are diagnosed with GI diseases each year. Gastroscopy detection is the only method for recognizing, early screening, and treatment method for the Upper GI tract; it plays a vital role in disease detection and treatment [2]. Although there are standardized processes for endoscopy and the report has template structures ("site + lesion + treatment") [3, 4], these processes still rely on the doctor's subjective evaluation and need manual input [5]. During the actual process, doctors need to continuously complete lesion observation, content description, and finish structured records under high pressure and intense environments [6]. Such a process would be easily affected by fatigue and limited by experience, which may lead to judgment errors and incomplete information.

In recent years, with the rise of multimodal large language models (MLLMs), visual-language collaborative AI systems have begun to emerge in the medical field [7, 8]. In the field of colonoscopy, there are already many achievements, such as DeFrame, that can be used as benchmarks for tasks of colorectal polyp recognition, lesion detection.[9] Works like ColonGPT is well-suited for image and text generation and classification [10]. However, multimodal research in the field of the GI system is still extremely lacking. The existing related research is mostly focused on single-modal image tasks; the combination of images and text and the establishment of a unified framework are still limited [11].

With those situations, our paper provides a model of multimodal tasks of upper gastrointestinal endoscopy for the first time. Based on the current advanced visual language architecture, we designed a unified task instruction fine-tuning process to support the unified processing for

multiple core tasks, including anatomical structure recognition, lesion diagnosis, lesion localization, and medical report generation assistants. It also integrates various public and clinical expert-annotated data sets, covering multiple GI Structures and lesion types such as the stomach, esophagus, and duodenum. The model framework we proposed not only has good generalization capabilities but also supports natural language interaction, which can assist doctors in intelligent interpretation and information exchange, improving efficiency while enhancing clinical practicality.

2. Related Work

In recent years, the application of AI in upper gastrointestinal endoscopy has made significant progress. Results such as the identification of early gastric cancer, Barrett's esophagus reach more than 90% accuracy [12]. However, current AI research in upper gastrointestinal endoscopy is still primarily focused on single-modal/image diagnosis tasks [13].

Large-scale multimodal instruction-tuning models such as ColonINST/ColonGPT have emerged in the field of lower gastrointestinal endoscopy, but the multimodal evaluation benchmark in the upper gastrointestinal tract remains largely unresolved. Specific related work is as follows:

2.1. Upper GI Related Works

The upper GI field remains dominated by single-modal image tasks. The majority of publicly published and widely cited AI papers in this area focus on image classification, segmentation, and detection (e.g., gastric cancer detection, site identification, and duodenal polyp localization) recently [11].

Upper GI AI research is dominated by image detection/classification: A review by Tokat et al. (2022) noted that AI applications in upper GI endoscopy primarily focus on tasks such as "identifying, characterizing, and localizing lesions in the upper GI tract," achieving accuracies up to 90% [4]. However, these applications remain limited to visual imagery and lack the integration of language or multimodal tasks [13].

Multimodal approaches are still in their infancy: Some applications have involved multimodal learning involving image and language, but this remains an experimental challenge and has yet to be widely applied [14]. Furthermore, it lags behind natural language interaction for medical diagnostic tasks. Some practical applications, such as EndoDINO, rely on purely image-based models (visual encoders) without language modules [15]. There have been a few attempts at medical report generation (e.g., using CNNs, LSTMs, or Transformers to generate diagnostic text), but these efforts typically focus on image-to-caption translation rather than structured multi-task, multi-modal instruction-tuning and multi-turn medical dialogue [16].

These results focus on classification or segmentation tasks such as anatomical part recognition and lesion detection. Many papers focus on the "image classification/segmentation/detection" task, lacking a unified multimodal, cross-task framework.

2.2. Previous Work of ColonGPT at Colonoscopy

In recent years, the previous work of ColonGPT has started to use multimodal large models for multi-task diagnosis. Examples include polyp detection, description generation, and decision support. These methods often train with image–text pairs and use instruction tuning to build a framework for understanding vision-language.

By using four downstream tasks, CLS, REC, REG, and CAP, these models align medical images with natural language. This lets the model switch between classification, localization, and description by only changing the language input, without changing the model structure. At the same time, the generative design improves the ability of the model to explain results to doctors, which makes it easier to use in clinical support systems [10]. However, large-scale multimodal instruction-tuning and vision–language dialogue, as seen in colonoscopy, have not yet been applied to the upper GI field. This is the main point of innovation in our study.

2.3. Multmodal Model and Instruction-Tuning

In recent years, multimodal foundation models have grown fast. LLaVA [17] and MiniGPT-4 [18] show that instruction-tuning can use joint vision and language instructions to unify many downstream task interfaces. This works well for general tasks such as image understanding, localization, image description, and multi-turn dialogue. In medicine, this method can unify work such as anatomical recognition, lesion diagnosis and localization, and report generation.

In the medical field, models like Med-Flamingo [19] and BioMedGPT [20] showed that multimodal methods can work in clinical tasks. But this kind of method is still rare in the upper GI endoscopy field. This is the starting point and main drive of our study. Works about models with proper size and new vision encoders can help. Phi-1.5 is a compact language model that learns from high-quality data and synthetic reasoning traces [21]. It shows strong reasoning ability with less parameters than larger models. SigLIP 2 improves the original SigLIP encoder with better alignment of text and image features [22]. It is efficient and supports high-resolution medical images. The combination of language models like these can give a good balance of accuracy, speed, and cost.

3. Proposed Approach

We design a unified multimodal framework for upper gastrointestinal (GI) endoscopy. The framework takes both image and text as input, and combines a vision encoder and a large language model (LLM). The vision encoder extracts semantic features from the endoscopic image, which are projected into the LLM embedding space through a mapping layer. The projected visual tokens are then fused with instruction text into a single sequence for multimodal reasoning.

3.1. Overall Framework Design

Our multimodal framework has three main parts: the vision encoder, the multimodal fusion layer, and the part of local LLM model. The vision encoder turns an endoscopy image into dense visual tokens. The fusion layer uses a linear projection to map the visual features to the language model embedding space. The local Multimodal model (such as LLaMA or Qwen) takes the fused multimodal input in a unified instruction–answer format and generates outputs for all tasks [23, 24]. As shown in Fig.1(GI frame from HyperKvasir[25]), this design allows all tasks to share the same input–output format and decoding logic. The method details are listed as follows:

1. **Vision Encoder**: We use a pretrained CLIP-ViT or LLaVA vision encoder to get high-level semantic features, like how previous work [10] get features from endoscopy images. The encoder turns the image into dense feature vectors (visual tokens) that keep both global context and local details. We use a large-scale pretrained vision network, such as SigLIP, or the vision tower of LLaVA. The input image is resized to a fixed size, split into patches, and encoded into a specific number(N_v) of visual tokens with position information. We can use attention pooling or learnable pooling to reduce the number of tokens to N_v' to lower computation cost and sequence length. We first use general large-scale pretraining and then do small-scale fine-tuning in the medical domain to speed up convergence and reduce GPU memory usage.
2. **Multimodal Fusion Layer**: We use a visual projection head (linear mapping) to map visual features into the token space of the language model and align the visual and text context. We use one or two linear layers to map the visual features from the original dimension to the hidden size of the language model. The visual tokens are then concatenated with text tokens into one sequence, so the language model can use self-attention for cross-modal interaction. This linear mapping has few parameters, works with existing LLM structures, and does not need major architectural changes.
3. **Large Language Model (LLM) Part**: We build a multimodal framework based on LLaVA as the language backbone [17]. In our framework, the LLM reads task instructions and processes with fused multimodal features from the vision encoder via a projection layer and lightweight adapters. During training, only the multimodal projection layer and adapter modules are updated, while the core LLM parameters are frozen or fine-tuned. We employ a unified conversational format (human: instruction + image; gpt: answer), so that different downstream tasks (CLS, REC, REG, CAP) follow the same input–output pattern, allowing the reuse of a single decoding logic across all tasks.
4. **Downstream tasks**: We follow the four downstream tasks: classification (CLS), localisation (REC), referring expression generation/area identification (REG), and image captioning (CAP). In our Upper GI multimodal framework, these four tasks match different diagnostic needs [10]. CLS classifies anatomy and lesion types to identify the category of site and disease. REC predicts the spatial location of the lesion. REG identifies the category in a target area. CAP generates medical descriptions of the endoscopic findings. This task setup gives a shared input–output format for multimodal fusion and instruction tuning, and builds the base for the fusion and training process.

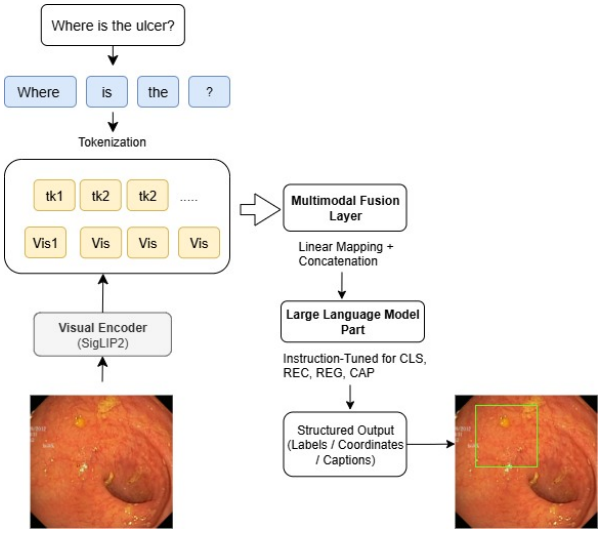

Figure 1. Overview of the Proposed Framework

3.2. Multimodal Fusion

After finalizing the framework design and completing the construction of downstream tasks, the model runs in the steps shown in Fig.2(GI frame from HyperKvasir[25]):

1. Visual feature extraction: We use the vision encoder to process upper GI endoscopy images. The image is normalized and resized, then split into fixed-size patches. Each patch is encoded into a visual token with position

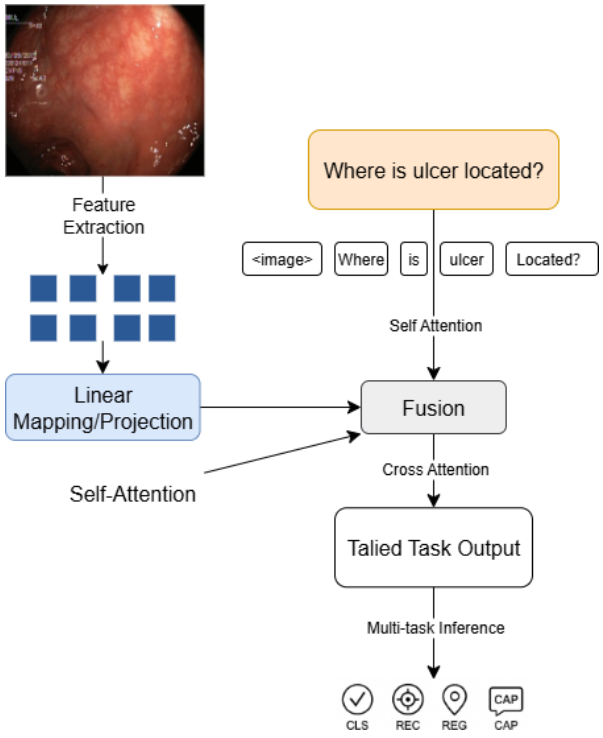

Figure 2. Multimodal feature fusion process

information, keeping both global context and local details.

2. Feature projection: Visual features from the encoder are mapped to the same hidden size as the language model by a linear mapping layer. This makes the visual tokens and text tokens share the same semantic space.
3. Task instruction input: We combine the projected visual tokens with the downstream task instruction into one input sequence in a unified multimodal format.
4. Multimodal feature fusion: Inserting a special image token to mark the place of visual information in the sequence. The language model uses self-attention to align and combine the visual and text features. Optionally, we can use cross-attention or feature concatenation in the fusion layer to improve alignment.
5. Multi-task inference: Using a unified instruction-tuning strategy, the local language model produces the output according to the task type. The output can be a structured class label, a recognition result, bounding box coordinates, or multi-sentence medical descriptions and supporting multi-turn natural language interaction with physicians.

3.3. Training Process

To make one model for all upper GI tasks, we use a two-stage training process with one instruction-tuning strategy. This lets the model do classification, recognition, localization, and description in one framework.

3.3.1. Stage 1: Image and Text Pre-alignment

We map the features from the vision encoder to the same semantic space as the language model. This makes visual tokens and text tokens share the same embedding size and meaning space. We use paired multimodal data (image + text description) for supervised training. We minimize the distance between image and text features in the embedding space so that visual and text content with the same meaning are close to each other. We use cross-entropy loss or contrastive loss to improve vision–text alignment.

3.3.2. Stage 2: Multi-task Joint Fine-tuning

We put all tasks (CLS, REC, REG, CAP) into one dialogue format as in previous work used [10]:

- Input (Human): task instruction + image path
- Output (GPT): task target (class label, bounding box, medical description, etc.)

All tasks share the same decoder and parameter space. The task type is decided by the instruction, so no separate model is needed. We mix batches from different tasks in training.

For each task, we use its metric: accuracy for CLS and REG, IoU for REC. With this unified instruction and multi-task mixed training, the model can switch tasks in clinical use without declaring the task type in advance. It just reads the natural language instruction and does the right task. This keeps vision–language alignment, allows multi-task unified reasoning, and supports natural language interaction. It gives the model more flexibility and practical value for clinical diagnosis support.

4. Experiments and Results

4.1. Dataset Collection and Construction

Compared with lower GI(Colonoscopy) datasets, upper GI datasets from public and private sources show big differences in labeling and task coverage. Upper GI detections cover multiple anatomical regions (esophagus, stomach, duodenum) and have more complex and variable lesions. This makes clinical collection and labeling harder to standardize. At present, no unified multimodal dataset exists that covers multiple tasks.

Split	CAP	CLS	REC	REG
train	64,349	31,000	16,704	16,645
val	7,242	3,635	1,899	1,890
test	15,225	10,048	2,585	2,579

Table 1. Data statistics (upper GI endoscopy images & dialogue pairs)

4.1.1. Dataset Collection

To fill this gap, we built the GastroINST dataset for our work. The dataset comes from two main sources: public datasets and private clinical datasets, with 64349 pictures. Table 1 shows the overall distribution of images across the training, validation, and test collections, with a total of 128,698 image–text pairs. The CLS task contains 31,000 training samples. The REC and REG tasks have almost the same number of training samples, with 16,704 and 16,645, respectively, showing a similar proportion.

Image–text pairs for the captioning (CAP) task were created by collecting images from all datasets, reconstructing them into JSON format, and combining the original text annotations with manually added descriptions. The image descriptions were created in a unified format by the local multimodal model mentioned in 3.2. Those multimodal models, such as LLaMA, can utilize prompt templates from previous work to produce professional image analysis and detailed descriptions after testing, ensuring that the final results have a consistent standard in a clinical environment.

Table 2 lists the details of the dataset, sources, types of related lesions, and the downstream task categories. All datasets were reformatted into a unified structure suitable for multi-task training based on their existing annotations. For example, GA-IM-Benchmark[29] and gastric_positions_new11_v1.3[26] contain positions and lesion category labels, most of which are from public sources and are used for CLS tasks. Datasets with lesion bounding boxes or segmentation masks, such as gastric_cancer_v2_test[27], Gastric Dataset[28] and kvasir-seg[31], come from public data and can support lesion recognition (REC) and localization (REG) tasks.

4.1.2. GastroINST Dataset Construction

Following the design principles of the ColonINST dataset and the clinical standards defined in the Kyoto Classification of Gastritis [33], we constructed the GastroINST dataset, whose hierarchical structure is illustrated in Fig. 3

Three-level taxonomy The labels are organized in three levels, covering the full scope from anatomy to pathology and special procedures, as shown in Table 3:

- Negative (no lesion) – Normal cases such as normal gastric body or normal duodenum.
- Anatomical Structure – Esophagus (upper/mid/lower segments, cardia), stomach (body, antrum, angle, fundus, pylorus), duodenum (bulb, descending part), and others.
- Pathological Findings – Esophageal lesions (inflammation, polyps, cancers), gastric lesions (gastritis, gastric ulcer, precancerous lesions/polyps, gastric cancer), and duodenal lesions (duodenal ulcer, polyps).

Unified multimodal format Following previous work, all data is converted into a unified instruction–image–output format, enabling joint training and inference across tasks. We also combine private upper GI data with public datasets to increase diversity and generalization.

4.2. Evaluation Metrics

Experiment uses several public and internal upper gastrointestinal (GI) endoscopy datasets. Each dataset is split into training, validation(Includes training dataset), and test sets by the official split or by an 8:1:1 ratio. For each task, we choose datasets that match the type of annotation.

For classification (CLS) and lesion recognition (REC) tasks, we use accuracy as the metric. Accuracy is the number of correct predictions divided by the total number of samples. We compare the model's prediction with the ground truth to see if it is correct. Table 2 shows the image and task distribution.

For lesion localization (REG) tasks, we use intersection over union (IoU) as the metric. IoU measures the overlap between the predicted box and the ground truth box. It is the ratio of the intersection area to the union area. A value close to 1 means the prediction matches the ground truth well. In this study, predictions with IoU less than 0.5 are counted as incorrect.

For image captioning (CAP) tasks, we check the model's ability to generate medical descriptions. We do not measure this task with numbers. Instead, we use manual review and case studies to check quality.

Finally, we compare our framework with previous work on the same tasks and settings. We compare accuracy and localization performance. This shows how well our framework works for upper GI multimodal tasks and its possible value in real clinical use.

4.3. Training and Evaluation Setup

We test the framework for upper GI multimodal tasks. Inspired by the two-stage multimodal training strategy in pervious work at ColonGPT, we adapt and extend the approach to upper gastrointestinal endoscopy tasks with two stages: cross-modal feature alignment and multi-task instruction tuning. Also, we changed and expanded the method for upper GI endoscopy.

In the cross-modal feature alignment stage, we used image–caption pairs (CAP task) to link vision and language features.In the multi-task instruction tuning stage, we used a unified input–output format for four tasks: classification (CLS), lesion recognition (REG), localization (REG), and caption generation (CAP). Unlike the original method, we updated the vision encoder and also opened some language model parameters to fit medical expressions and task needs.

In the end, we measured performance on a validation set (Seen) and a test set (Unseen). For evaluation, we tested

Category	CLS	REC	REG	CAP	Total	Type	Source
gastric_positions_new_v11_v1_3 [26]	27677	N/A	N/A	N/A	27677	AS	Private
gastric_cancer_v2_train [27]	N/A	7750	17550	16030	55530	PF-ED	Private
Gastric Dataset [28]	N/A	7225	7451	14376	27852	PF-GD	Private
QA-IM-Benchmark [29]	11672	N/A	N/A	3523	15195	PF-ED	Private
esophageal_cancer_v2.0 [30]	N/A	5170	5170	10040	20680	PF-GD	Private
kvasir-dataset-v2 [31]	30000	N/A	N/A	N/A	30000	PF-GD	Public
GastroVision [32]	2334	N/A	N/A	2347	4681	PF-ED	Public
gastric_cancer_v2_test [27]	N/A	1043	1043	2086	4172	PF-GD	Private

Table 2. Dataset categorization — mapping each dataset to three-level classification scheme. Abbreviations: AS = Anatomical Structure, PF-ED = Pathological Findings (Esophageal Diseases), PF-GD = Pathological Findings (Gastric Diseases) as shown in Table 3 and Fig. 3

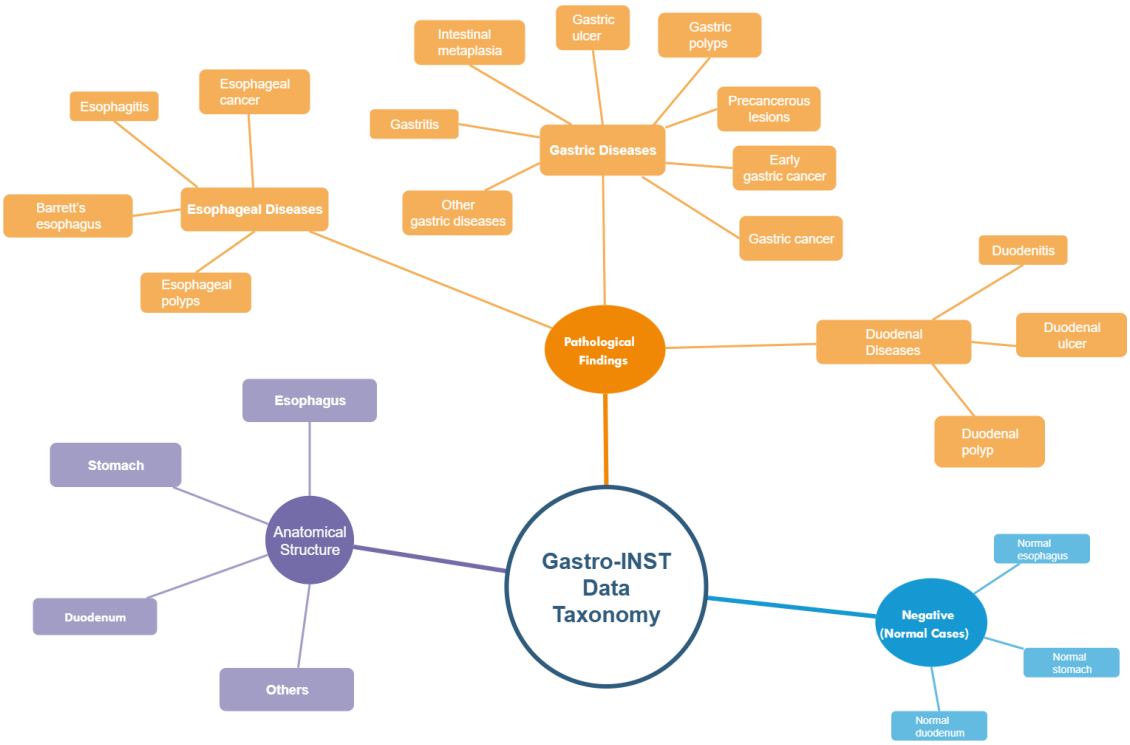

Figure 3. Gastro-INST Data Taxonomy

CLS, REC, and REG on Seen and Unseen sets. CLS and REG used accuracy and REC used IOU with default threshold(0.5) for area match for checking correctness.

5. Results Analysis and Evaluation

5.1. Model Performance Results and Analysis

For the overall results from table 4, the CLS task kept high accuracy in both Seen and Unseen sets (91.36% / 88.63% vs ColonGPT 94.06% / 83.24%). This shows our framework has better cross-domain generalization for upper GI classification. The REG task was stable in Seen and Unseen (77.57% / 76.89% vs ColonGPT 85.74% / 56.24%). The drop from Seen to Unseen was small, which means the localization ability is more stable across different data. The REC task was much lower than CLS and REG (38.81% / 37.41% vs ColonGPT 85.74% / 56.24%), which is the main limitation. The gap may be caused by high shape variation of lesions, small sample size, and unbalanced class distribution in upper GI data.

In the detailed results by table 5 and Table 6, for CLS, large datasets such as GA-IM-Benchmark and and gastric_positions_new11_v1.3 had accuracy close to or over 90%. GastroVision had lower accuracy (Seen 81.20%, Unseen 74.25%) because it is small and has high diversity. For REG, the results were stable (Seen and Unseen both in the 76–78% range). In esophageal_cancer2.0[30], Seen reached 83.25%, showing high consistency in lesion localization.

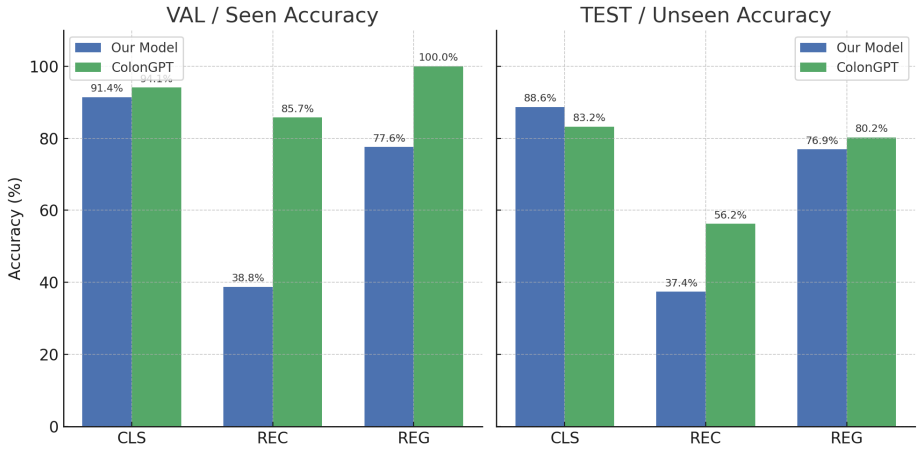

Figure 4. Accuracy Overview

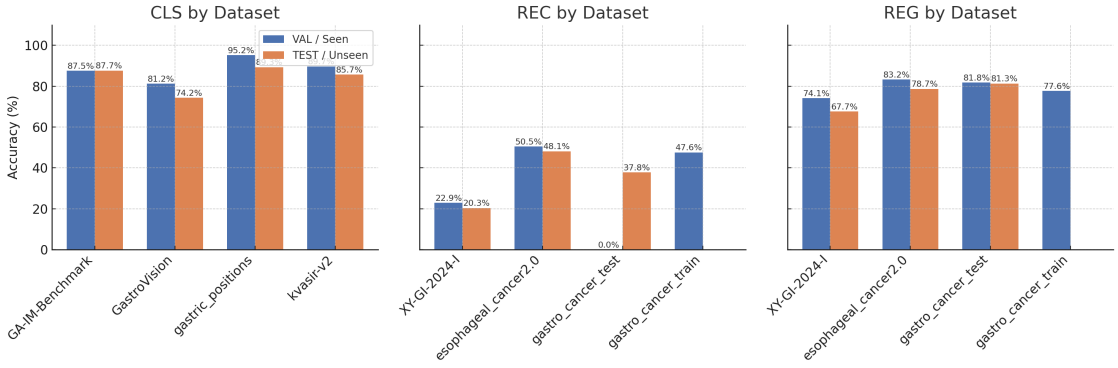

Figure 5. Accuracy View by DataSet

For REC part, the private data sets, especially the gastro cancer data sets, had the weakest results, probably due to complex lesion types and few samples. The esophageal dataset was better (about 50% in both Seen and Unseen), showing better learnability for some diseases.

5.2. Conclusion

The model reached high accuracy in key Upper GI tasks, such as common anatomy recognition and typical lesion classification. This shows good effectiveness for gastroscopy tasks. In complex tasks, the model improved its generalization ability. It can handle classification, localization, and medical description generation at the same time. For classes with enough samples and consistent labels, the model performed close to experts.

5.3. Limitations and Challenges

However, for some lesions and rare locations where data is scarce and class distribution is very unbalanced, the model's accuracy and localization dropped clearly for the REC tasks. The current data size and quality still limit generalization and detection ability. Medical description generation (caption) is usable, but the output format needs more control to match the standard report style used by physicians. Data size and quality limits also reduced performance in rare cases and with different device types. Generalization in these cases still needs work.

5.4. Overall Review

In conclusion, the results show that the proposed multimodal model has strong overall ability and stability in handling complex and diverse upper GI endoscopy tasks. It achieved good performance in anatomical structure recognition, typical lesion classification, and medical description generation. With performance close to expert doctors in some tasks, the model shows promise for future clinical use and deployment, and lays a solid foundation for building intelligent and automated upper GI diagnosis and treatment systems.

Level I	Level II	Level III (examples)
Negative	Normal	Normal esophagus, stomach, duodenum
Anatomical Structure	Esophagus	Major subregions
	Stomach	Major subregions
	Duodenum	Major subregions
	Others	Other sites
Pathological Findings	Esophageal Diseases	Typical lesions, cancer
	Gastric Diseases	Gastritis, ulcer, cancer
	Duodenal Diseases	Duodenitis, ulcer

Table 3. Three-level classification scheme for upper gastrointestinal endoscopy (referencing the Kyoto Classification of Gastritis [33]).

Task	Our Model(VAL)	Our Model(TEST)	ColonGPT (VAL)	ColonGPT (TEST)
CLS	91.36%	88.63%	94.06%	83.24%
REC	38.81%	37.41%	85.74%	56.24%
REG	77.57%	76.89%	99.96%	80.18%

Table 4. Overall Accuracy Comparison (VAL / TEST, with ColonGPT)

Task	Category	Accuracy (%)	Total
CLS	GA-IM-Benchmark	87.50	1168
CLS	GastroVision	81.20	234
CLS	gastric_positions_new11_v1.3	95.19	1933
CLS	kvasiar-dataset-v2	89.67	300
CLS	–	**91.36**	**3635**
REC	Gastric Dataset	22.86	700
REC	esophageal_cancer2.0	50.49	406
REC	gastric_cancer_v2_test	0.00	11
REC	gastric_cancer_v2_train	47.57	782
REC	–	**38.81**	**1899**
REG	Gastric Dataset	74.10	691
REG	esophageal_cancer2.0	83.25	406
REG	gastric_cancer_v2_test	81.82	11
REG	gastric_cancer_v2_train	77.62	782
REG	–	**77.51**	**1890**

Table 5. Accuracy Comparison by Dataset - Val/Seen

Task	Category	Accuracy (%)	Total
CLS	GA-IM-Benchmark	87.66	1167
CLS	GastroVision	74.25	233
CLS	gastric_positions_new11_v1.3	98.20	8348
CLS	kvasir-dataset-v2	85.67	300
CLS	–	**88.62**	**10048**
REC	Gastric Dataset	20.34	649
REC	esophageal_cancer2.0	48.15	999
REC	gastric_cancer_v2_test	37.78	937
REC	–	**37.41**	**2585**
REG	Gastric Dataset	67.55	643
REG	esophageal_cancer2.0	78.68	999
REG	gastric_cancer_v2_test	81.32	937
REG	–	**76.89**	**2579**

Table 6. Accuracy Comparison by Dataset Test/Unseen

5.5. Future Works

We plan to expand and balance the dataset. Tasks and categories still have limited data and an unbalanced distribution. We will collect more types of upper GI endoscopy images to increase overall size and balance, and reduce the performance gap caused by data shortage.

We will introduce more medical-specific large language models. We will also integrate specialized models for medical imaging, such as LingShu [34] and LLama-Med [8], and explore model fusion or distillation between general vision-language models and medical-specific models, aiming to maintain both medical expertise and cross-task generalization.

We will build a verification and rating system for output. Inspired by other work, we will add multi-metric and multi-dimensional evaluation during fine-tuning and testing, beyond only accuracy and IoU. For text generation and report tasks, we will use standard metrics such as BLEU-4 [35], ROUGE-1/2 [36], and BERTScore [37]. We will also explore using GPT-4 Judge to check semantic and factual consistency [38], combined with end-to-end expert scoring. This process can run regularly during training and fine-tuning to ensure that all outputs match clinical expectations.

We will fine-tune with structured report data and generate structured reports. Using the structure of professional endoscopy reports, we will feed structured diagnostic information, such as location, lesion type, treatment, and conclusion, with image features into the model for fine-tuning. With the fine-tuning of report data, the model can produce automated reports that follow hospital standards and can be used directly in clinical work.

References

[1] Anne F. Peery, Evan S. Dellon, Jennifer Lund, Seth D. Crockett, Christopher E. McGowan, William J. Bulsiewicz, Lisa M. Gangarosa, Michelle T. Thiny, Karyn Stizenberg, Douglas R. Morgan, Yehuda Ringel, Hannah P. Kim, Marco Dacosta DiBonaventura, Charlotte F. Carroll, Jeffery K. Allen, Suzanne F. Cook, Robert S. Sandler, Michael D. Kappelman, and Nicholas J. Shaheen. Burden of gastrointestinal disease in the united states: 2012 update. *Gastroenterology*, 143(5):1179–1187.e3, 2012. 1

[2] Aaron P. Thrift and Hashem B. El-Serag. Burden of gastric cancer. *Clinical Gastroenterology and Hepatology*, 18(3):534–542, 2020. 1

[3] Daphnée Beaulieu, Alan N Barkun, Catherine Dubé, Jill Tinmouth, Pierre Hallé, and Myriam Martel. Endoscopy reporting standards. *Can J Gastroenterol*, 27(5):286–292, 2013. 1

[4] Meltem Tokat, Laurelle van Tilburg, Arjun D Koch, and Manon C W Spaander. Artificial intelligence in upper gastrointestinal endoscopy. *Dig Dis*, 40(4):395–408, July 2021. 1, 2

[5] Michael Bretthauer, Lars Aabakken, Evelien Dekker, Michal F Kaminski, Thomas Rösch, Rolf Hultcrantz, Stepan Suchanek, Rodrigo Jover, Ernst J Kuipers, Raf Bisschops, Cristiano Spada, Roland Valori, Dirk Domagk, Colin Rees, Matthew D Rutter, and ESGE Quality Improvement Committee. Reporting systems in gastrointestinal endoscopy: Requirements and standards facilitating quality improvement: European society of gastrointestinal endoscopy position statement. *United European Gastroenterol J*, 4(2):172–176, February 2016. 1

[6] Wenju Du, Nini Rao, Dingyun Liu, Hongxiu Jiang, Chengsi Luo, Zhengwen Li, Tao Gan, and Bing Zeng. Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. *IEEE Access*, 7:142053–142069, 2019. 1

[7] Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey, 2025. 1

[8] 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, 2023. 1, 8

[9] Shuijiao Chen, Shuang Lu, Yingxin Tang, Dechun Wang, Xinzi Sun, Jun Yi, Benyuan Liu, Yu Cao, Yongheng Chen, and Xiaowei Liu. A machine learning-based system for real-time polyp detection (deframe): A retrospective study. *Frontiers in Medicine*, Volume 9 - 2022, 2022. 1

[10] Ge-Peng Ji, Jingyi Liu, Peng Xu, Nick Barnes, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. Frontiers in intelligent colonoscopy, 2025. 1, 2, 3, 4

[11] Ziyi Jin, Tianyuan Gan, Peng Wang, Zuoming Fu, Chongan Zhang, Qinglai Yan, Xueyong Zheng, Xiao Liang, and Xuesong Ye. Deep learning for gastroscopic images: computer-aided techniques for clinicians. *BioMedical Engineering OnLine*, 21(1):12, Feb 2022. 1, 2

[12] Meltem Tokat, Laurelle van Tilburg, Arjun D Koch, and Manon C W Spaander. Artificial intelligence in upper gastrointestinal endoscopy. *Dig Dis*, 40(4):395–408, July 2021. 2

[13] Sharib Ali. Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions. *npj Digital Medicine*, 5(1):184, Dec 2022. 2

[14] Triet M. Thai, Anh T. Vo, Hao K. Tieu, Linh N. P. Bui, and Thien T. B. Nguyen. Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering, 2023. 2

[15] Patrick Dermyer, Angad Kalra, and Matt Schwartz. Endodino: A foundation model for gi endoscopy, 2025. 2

[16] E. Gong, C. Bang, J. Lee, J. Park, E. Kim, S. Kim, M. Kimm, and S. Choi. Large language models in gastroenterology: Systematic review. *Journal of Medical Internet Research*, 26(1):e66648, 2024. 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 3

[18] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2

[19] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023. 2

[20] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, August 2024. 2

[21] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023. 2

[22] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2

[23] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, et al. The llama 3 herd of models, 2024. 3

[24] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3

[25] Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K. Stensland, Enrique Garcia-Ceja, Peter T. Schmidt, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, and Thomas de Lange. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, Aug 2020. 3

[26] Gastric Positions Dataset V1.3 (gastric_positions_new11_v1.3), 2024. Private internal dataset for anatomical classification. Task: CLS (anatomical classification). Contains 11 gastric positions. Annotation format: YOLO. Not publicly available. 5, 6

[27] Gastro Cancer Dataset, 2024. Private internal dataset. Task: REC (lesion recognition). Focus on high-risk gastric lesions. Not publicly available. 5, 6

[28] Gastric Dataset, 2024. Internal dataset. Task: REC (lesion recognition). Lesion types include gastritis, gastric ulcer, gastric polyp, gastric cancer, precancerous lesions (low/high grade), and others. Annotation format: COCO. Not publicly available. 5, 6

[29] Jie Yang, Yan Ou, Zhiqian Chen, Juan Liao, Wenjian Sun, Yang Luo, and Chunbo Luo. A benchmark dataset of endoscopic images and novel deep learning method to detect intestinal metaplasia and gastritis atrophy. *IEEE J Biomed Health Inform*, 27(1):7–16, January 2023. 5, 6

[30] Esophageal Cancer 2.0 Dataset, 2024. Private internal dataset. Task: REC (lesion recognition). Contains esophageal cancer samples. Not publicly available. 6

[31] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pal Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, 2017. 5, 6

[32] Debesh Jha, Vanshali Sharma, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, M. K. Bhuyan, Pradip K. Das, Michael A. Riegler, Pål Halvorsen, Ulas Bagci, and Thomas de Lange. Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection, 2023. 6

[33] Ken Haruma. *Kyoto Classification of Gastritis ()*. Nankodo, Tokyo, 3rd ed. edition, 2023. Originally published in Japanese as " (Ien no Kyoto Bunrui)". 5, 8

[34] LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning, 2025. 8

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. 8

[36] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 8

[37] Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. CIDEr-R: Robust consensus-based image description evaluation. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online, November 2021. Association for Computational Linguistics. 8

[38] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 8