A Clinician-Guided Framework for Endoscopic AI: Developing PanEndoAtlas and Benchmarking Foundation Models Across the Full GI Spectrum

Shreya Johri[1], Luyang Luo PhD[1], Hong-Yu Zhou PhD[1], Todd Brenner MD[2], Sami Elamin MD[2], Mark Enrik Geissler MD[2], Tyler M. Berzin MD[2]*, Pranav Rajpurkar PhD[1]*

1. Department of Biomedical Informatics, Harvard Medical School
2. Center for Advanced Endoscopy, Beth Israel Deaconess Medical Center and Harvard Medical School

Endoscopic procedures play a central role in the diagnosis and management of gastrointestinal (GI) diseases, yet the field lacks large-scale, clinically diverse benchmarks and unified datasets to evaluate vision foundation models. We introduce PanEndoSuite, the first unified ecosystem for endoscopic AI, developed through systematic collaboration between AI researchers and practicing gastroenterologists. PanEndoSuite consists of three complementary components: PanEndoAtlas, PanEndoX, and PanEndoFM. PanEndoAtlas is a harmonized dataset of over 420,000 labeled images from 30 public endoscopy datasets across 13 countries and 26 hospitals, creating a clinically-grounded hierarchical taxonomy that mirrors diagnostic reasoning patterns across 111 GI diseases. PanEndoX is a benchmark of 10 clinically grounded tasks, including hierarchical GI-tree classification, Barrett's esophagus grading, ulcerative colitis scoring, polyp subtyping, Boston Bowel Preparation Scale assessment, multi-organ disease classification, and anatomical landmark identification—designed to probe generalization across anatomical regions, disease presentations, and annotation granularities. PanEndoFM is a foundation model pretrained on a 10 million–image corpus curated from public data sources, spanning the entire GI tract. We benchmark PanEndoFM against two endoscopy-specific foundation models (EndoFM-LV, EndoSSL) and two general-purpose vision models (ViT-B/16, ResNet-50). PanEndoFM achieves the highest macro-AUC on 6 of 10 tasks, demonstrating broad clinical generalization; EndoFM-LV performs best on colon-focused tasks, EndoSSL excels in polyp subtyping, and ViT-B/16 shows strengths on small-intestine conditions. Together, PanEndoSuite establishes a foundation for building robust, generalist AI systems in gastrointestinal endoscopy that bridge current AI capabilities and clinical practice.

Keywords: Endoscopic Diagnosis; Vision Foundation Models; Datasets and Benchmark; PanEndoSuite; PanEndoAtlas; PanEndoFM

1. Introduction

Clinical integration of foundation models requires rigorous evaluation for real-world clinical use cases. Large and diverse datasets, along with robust benchmarks featuring clinically meaningful tasks, are essential to ensure this alignment. In domains like pathology and radiology, large-scale vision foundation models have demonstrated strong generalization with minimal

*Corresponding authors: pranav_rajpurkar@hms.harvard.edu, tberzin@bidmc.harvard.edu

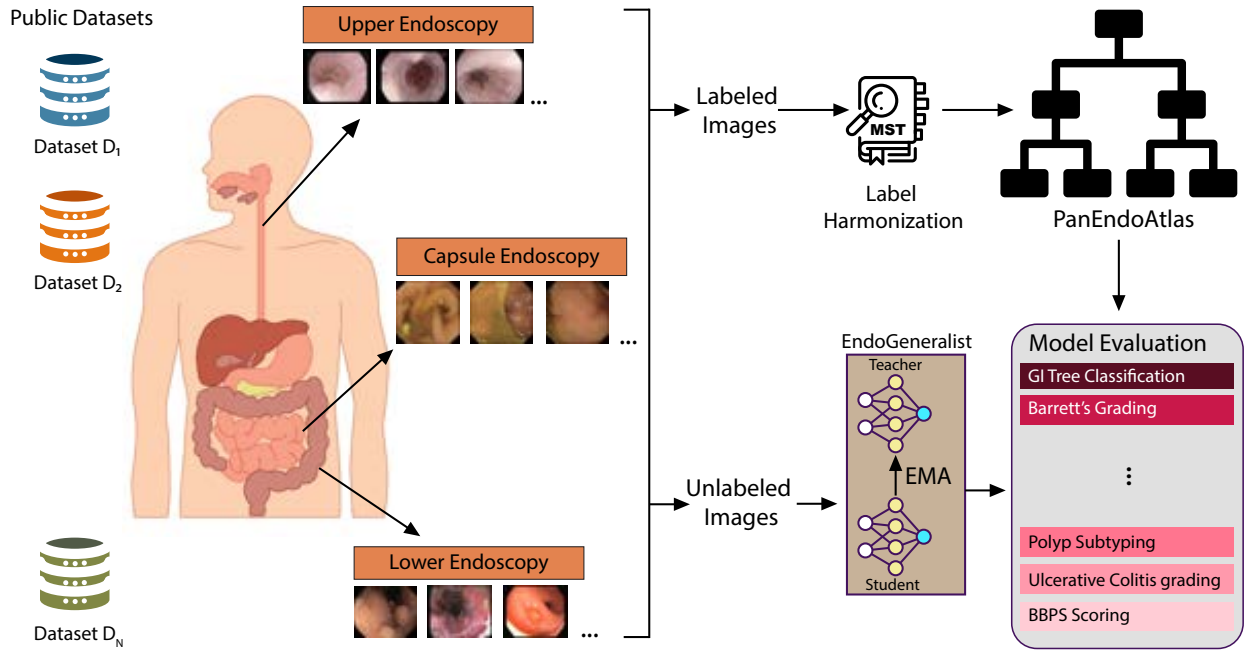

Fig. 1: PanEndoAtlas and PanEndoX Overview. PanEndoAtlas curates more than 420,000 labeled images spanning 111 GI abnormalities, from 13 countries and over 26 hospitals. PanEndoX is a benchmark suite of 10 clinically important GI tasks.

fine-tuning.[1–3] This progress has been accelerated by curated datasets and benchmarks designed to evaluate generalizability, robustness, and foundational capability.[4,5]

In contrast, gastroenterology lacks unified benchmarks despite the central role of endoscopy in diagnosing gastrointestinal (GI) diseases, including early cancer detection.[6,7] Diagnostic accuracy is often hampered by subtle disease presentations and clinician fatigue during high-throughput procedures.[8,9] Although several open-source and commercial vision models have been proposed for endoscopy,[10–14] their general-purpose utility remains unproven. Existing public datasets are fragmented, narrowly focused on common conditions such as colon polyps, and lack standardized annotations,[15–17] limiting clinical relevance and model generalization.

To address these limitations, we introduce PanEndoSuite, the first unified ecosystem for endoscopic AI, developed in collaboration between AI researchers and gastroenterologists. PanEndoSuite consists of three complementary components: PanEndoAtlas, PanEndoX and PanEndoFM. PanEndoAtlas is a harmonized dataset of more than 420,000 labeled images across 111 GI abnormalities, curated from 30 public datasets spanning 13 countries and 26 hospitals, with standardized hierarchical labels based on the MST-3.0 framework.[18] PanEndoX is a benchmark of 10 clinically grounded tasks including clinical disease classification, severity grading, and anatomical localization, to evaluate the foundational capabilities of endoscopic vision models. PanEndoFM is a vision foundation model pretrained via DINOv2[19] methodology on 10 million endoscopic images.

We benchmark PanEndoFM against two endoscopy-specific foundation models (EndoFM-LV, EndoSSL) and two general-purpose vision models (ViT-B/16, ResNet). PanEndoFM

achieves the highest AUC on 6 out of 10 benchmark tasks, demonstrating strong generalization across clinical scenarios. Additionally, we observe domain-specific strengths in other models, highlighting the influence of pretraining data on downstream performance. Together, these contributions establish PanEndoSuite as a foundation for building robust, generalist AI systems in gastrointestinal endoscopy that bridge current AI capabilities and clinical practice.

2. Related Work

Endoscopy Datasets: Public endoscopy datasets are fragmented and institution-specific, limiting their generalizability. Early datasets consisted of only a few hundred labeled images,[20,21] and while more recent efforts have scaled to several thousand examples,[22] they remain isolated contributions. Since 2012, over 25 public datasets have been released from more than 13 countries,[23] yet these are typically built independently, using different imaging protocols and patient populations, which hinders efforts to create unified training or evaluation sets.

Labeling across datasets is also highly inconsistent. While some groups maintain internal consistency within their own releases,[24,25] the broader landscape lacks a standardized annotation vocabulary. As a result, conceptually similar findings are labeled differently across datasets, or the same label is used for different anatomical contexts. For example, the label "polyp" may refer to lesions found in the colon, stomach, small intestine, or esophagus, creating semantic ambiguity. These inconsistencies make it difficult to merge datasets or perform cross-dataset evaluation without extensive re-labeling and clinical review.

Finally, current datasets are clinically narrow, with a strong bias toward colorectal polyps. Despite endoscopy being routinely used to diagnose and monitor over 100 distinct gastrointestinal conditions, most public datasets focus almost exclusively on polyp detection or segmentation.[15,16,26] A few recent datasets attempt to broaden disease coverage,[22] but they tend to be highly imbalanced, with a long tail of underrepresented conditions that models struggle to learn from. This limits the clinical relevance and diagnostic scope of models trained on existing resources.

To overcome these limitations, we introduce a large-scale benchmark dataset that is standardized, clinically diverse, and built using a hierarchical labeling system aligned with real-world endoscopic practice.

Endoscopy Foundation Models: A growing number of endoscopy foundation models have been developed using diverse pretraining strategies and datasets, yet their evaluations remain narrow in clinical scope. EndoSSL[10] used a masked siamese network pretrained on 2 million colonoscopy frames enriched for polyp-containing images, and was evaluated on polyp detection and classification tasks using single-center datasets focused exclusively on colon-polyps. EndoFM[11] scaled pretraining to 33,000 colonoscopy videos (5 million frames) and adopted a video transformer to capture spatiotemporal features for lesion detection and segmentation. Its variant, EndoFM-LV,[12] further extended training to over 12 million frames for longer video sequences. Both EndoFM and EndoFM-LV have been evaluated on tasks involving colon-polyp classification, detection, and segmentation. EndoMamba[13] adopted a state-space model architecture[42] in place of transformers,[43] and trained on combined 8 million

Table 1: Datasets encompassing PanEndoAtlas.

S.No.	Year	Dataset	Country	Hospital
1.	2012	ETIS-Larib[20]	France	Lariboisière Hospital-APHP
2.	2015	CVC-ClinicDB[21]	Spain	Hospital Clinic Barcelona
3.	2016	Kvasir[24]	Norway	Bærum Hospital
4.	2017	Nerthus[27]	Norway	Bærum Hospital
5.	2019	Kvasir-SEG[28]	Norway	Bærum Hospital
6.	2020	EDD2020[29]	France, Italy, UK	Ambroise Pare Hospital of Boulogne-Billancourt, Paris, France; Centro Riferimento Oncologico IRCCS, Aviano, Italy; Istituto Oncologico Veneto, Padova, Italy; John Radcliffe Hospital, Oxford, UK
7.	2020	HyperKvasir[25]	Norway	Bærum Hospital
8.	2020	CP-CHILD[30]	China	Hunan Children's Hospital
9.	2021	KvasirCapsule[31]	Norway	Bærum Hospital
10.	2021	KvasirCapsule-SEG[32]	Norway	Bærum Hospital
11.	2021	Kvasir-Sessile[33]	Norway	Bærum Hospital
12.	2021	KUMC[34]	US	University of Kansas Medical Center
13.	2021	SUN[35]	Japan	Showa University Northern Yokohama Hospital
14.	2021	CrohnIPI[36]	France	Nantes University Hospital
15.	2022	SUN-SEG[37]	Japan	Showa University Northern Yokohama Hospital
16.	2022	ERS[22]	Poland	Gdańsk University of Technology
17.	2023	PolypGen[15]	France, Italy, Norway, UK, Egypt	Ambroise Paré Hospital, Paris; Istituto Oncologico Veneto, Padova, Italy; Centro Riferimento Oncologico, IRCCS, Italy; Oslo University Hospital, Oslo; John Radcliffe Hospital, Oxford; University of Alexandria, Alexandria
18.	2023	GastroVision[38]	Norway, Sweden	Bærum Hospital, Karolinska University Hospital
19.	2023	MedFMC-Endo[39]	China	Renji Hospital
20.	2023	AICE[40]	Japan	Kyushu University Hospital
21.	2023	POLAR[41]	Netherlands, Spain	8 hospitals

still images and 10,000 videos. Its evaluations, like other models, have focused on colon-polyp tasks. Finally, EndoDINO,[14] a commercial model, trained DINOv2 on 10 million images, selected from a larger pool of 3.5 billion video frames, and performed validation on individual datasets. However, due to the unavailability of their model weights publicly, independent validation is limited.

Evaluation protocols across these models vary widely and remain constrained to colon-polyp related tasks and single-center datasets. The lack of standardization in validation data

and task diversity makes it difficult to perform fair comparisons or assess model generalization across different clinical settings and gastrointestinal conditions.

To address these gaps, we introduce a multi-center benchmark dataset that spans a broader spectrum of GI diseases and enables consistent, comprehensive evaluation of endoscopy foundation models across diverse diagnostic tasks.

3. Dataset

3.1. *PanEndoAtlas*

We introduce PanEndoAtlas which is a multi-source benchmark dataset designed to support standardized evaluation of AI models across a wide range of gastrointestinal endoscopy tasks. It comprises over 420,000 labeled images drawn from 21 public datasets, representing 13 countries and more than 26 clinical institutions (Table 1, Figure 2). The dataset spans all major anatomical regions of the GI tract, including the esophagus, stomach, small intestine, and colon, and includes a range of visual content, such as normal anatomy, pathological findings, therapeutic interventions, and quality control annotations.

Fig. 2: World map showing geographic distribution of PanEndoAtlas' constituent datasets across 13 countries and more than 26 clinical institutions. The dataset spans all major anatomical regions of the GI tract.

To unify the diverse label spaces in the constituent datasets, all annotations were systematically mapped into a standardized vocabulary inspired by MST-3.0 taxonomy.[18] These labels were further organized into a clinically-relevant four-level hierarchy to support training and evaluation of models on rare or sparsely represented categories (Figure 3).

- **Level 1: Anatomical Region.** Each image was assigned to one of four regions, upper GI (esophagus, stomach) or lower GI (small intestine, colon), based on available meta-

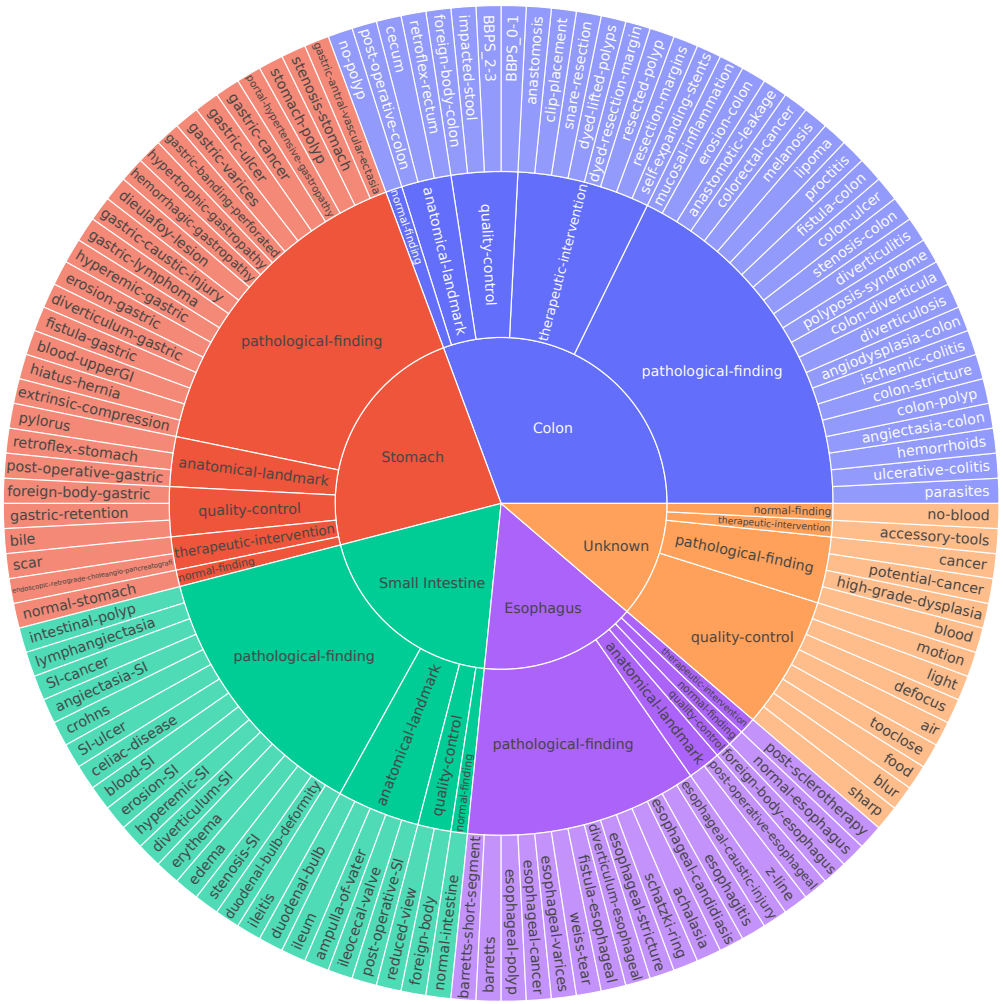

Fig. 3: PanEndoAtlas.

data, associated publications, or the type of imaging instrument used. Images with no definitive anatomical metadata were labeled as 'unknown'.

- **Level 2: Broad Clinical Category.** Images were categorized into one of five clinically relevant types: normal findings, anatomical landmarks, pathological findings, therapeutic procedures, and quality control measures.
- **Level 3: Fine-Grained Clinical Labels.** This level provides more specific diagnoses or findings within each broad category. Examples include pathological conditions such as polyps, ulcerative colitis, and Barrett's esophagus; anatomical landmarks such as the z-line, pylorus, and cecum; and procedural annotations such as polyp resection.
- **Level 4: Subtype and Severity Annotations.** Where available, additional clinical detail was included. For example, polyps were subclassified into adenomatous, hyperplastic, serrated, sessile and invasive-cancer. Ulcerative colitis cases included MAYO severity scores, and Barrett's esophagus cases were labeled according to subtype (e.g., short-segment, high-grade-dysplasia).

3.2. *Label Harmonization*

Each individual dataset in PanEndoAtlas had its own mapping dictionary, created in collaboration with a Gastroenterology Fellow. These mappings aligned original dataset labels with the standardized PanEndoAtlas schema, using three main sources: the original labels, dataset descriptions from the associated publications, and documentation from the datasets' GitHub repositories. During this process, two major types of ambiguity were resolved: location ambiguity, where anatomical location was absent from the label (e.g., 'polyp', 'ulcer'), and medical synonym ambiguity, where different terms referred to the same clinical condition or finding (e.g., 'angiodysplasia' and 'angiectasia'; 'diverticulosis', 'diverticulum', and 'diverticula'; 'edema', 'erosion', 'hyperemic', 'erythema', and 'mucosal inflammation').

Level-1 (Anatomical Region): Inferred using endoscope type and clinical label. Ambiguous cases (e.g., 'potential-cancer' in a study containing images from multiple GI regions) were labeled 'unknown'. All images have a level-1 label.

Level-2 (Broad Clinical Category): Derived from fine-grained labels or taken directly from existing annotations. All images have a valid level-2 label.

Level-3 (Fine-Grained Clinical Label): Harmonization occurred in two stages. First, labels were made anatomically specific (e.g., 'polyp' → 'colon-polyp', 'stomach-polyp' etc.). Second, clinically guided combinations were applied (e.g., BBPS-0 and BBPS-1 were combined into 'BBPS-0-1' since both indicate inadequate bowel preparation clinically). Distinctions were preserved when clinically meaningful e.g., 'colorectal-cancer' and 'colon-polyp' were merged for intramucosal cancers but kept separate for invasive lesions. All images have a level-3 label.

Level-4 (Subtype/Severity): Present in only some datasets; harmonized to standard terminology. Multiple grading systems were retained where relevant. Missing annotations were labeled 'None'. Tasks for subtype/severity were built from the subset with valid level-4 labels.

3.3. *Pretraining Dataset*

We assembled a pretraining corpus of 10 million unlabeled endoscopic images for training PanEndoFM. These images were sourced from five public datasets: ColonoscopicDS[44] , LDPolypVideo[26] , ERS[22] , RI-VCE[45] , and EndoFM-pretrain[11] . These span the full GI tract, including the esophagus, stomach, small intestine, and colon. It incorporates images captured using a variety of endoscopic instruments, such as colonoscopes, gastroscopes, and capsule endoscopes. The data reflect clinical and geographical diversity, with contributions from 5 countries, and capture variation in anatomical presentation, imaging protocols, disease states, and bowel preparation quality.

4. Methods

4.1. *PanEndoFM Pretraining*

We used a ViT-L architecture as the backbone for PanEndoFM, initialized with the official DINOv2 weights.[19] We then continued pretraining using DINOv2 methodology on the 10 million-image corpus described above. Model performance during pretraining was monitored

using Top-5 K-nearest neighbors (KNN) accuracy, computed over 100 neighbors using the Kvasir dataset,[24] which contains eight equally balanced classes. The best checkpoint was selected based on this validation metric. Training was conducted for two epochs, after which early stopping was triggered due to convergence in KNN accuracy. All other hyper-parameters were kept same as default.

4.2. *PanEndoX Benchmark*

We benchmark the performance of PanEndoFM against four additional models: 2 existing open-source endoscopy foundation models (EndoSSL, EndoFM-LV) and 2 general-purpose vision models (ViT-B/16, ResNet-50). We used linear probing, where the backbone was frozen and only a classification head was trained. This enabled a true assessment of image encoding generated by different vision encoders. A set of clinically relevant gastrointestinal (GI) tasks were used to assess the generalization ability (Table 2).

Baseline Methods: We adapted EndoFM-LV[12] and EndoSSL[10] foundation models for image-level evaluation. We further implemented the ViT-B/16 and ResNet-50 model architectures, initialized with ImageNet pretrained weights.

Dataset Splits: For each task, the data was split into training, validation, and test sets in a 75:10:15 ratio to support reproducibility. Splits were first performed at the dataset level, ensuring that all images from a given patient appeared in only one of the three sets. In cases where patient identifiers were not available, samples were randomly assigned. Within each dataset, label proportions were preserved across the splits to maintain class balance. Finally, the splits from individual datasets contributing to a given task were merged into a single unified file.

Hyper-parameter Search: All models, including baseline and comparison models, were fine-tuned using a hyperparameter search over learning rates $[1e^{-4}, 5e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}]$. The best model was selected based on validation AUC, and performance on independent test set is reported. Optimization was performed using AdamW (weight decay = 0.01) with a cosine learning rate scheduler ($\eta_{\min} = 1e^{-6}$, $T_{\text{mult}} = 2$). Other fine-tuning settings were held constant across experiments: batch size = 128, and early stopping after 10 epochs without improvement on the validation set.

Loss Function: We use binary cross-entropy (BCE) loss for our multi-label classification setup:

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right], \tag{1}$$

where $y \in \{0, 1\}$ is the ground-truth label and $\hat{y} \in [0, 1]$ is the predicted probability.

To address class imbalance, the BCE loss is weighted using a log-scaled class prevalence term. The weight for class i is computed as:

$$w_i = \log\left(1 + \frac{N}{n_i}\right), \tag{2}$$

where N is the total number of samples and n_i is the number of samples belonging to class i. The resulting weighted BCE loss is:

$$\mathcal{L}(y, \hat{y}) = -w\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right], \tag{3}$$

Table 2: Task descriptions for PanEndoX Benchmark.

Task Name	**Description**
GI-Tree Classification	Hierarchical classification of GI anatomy, visual label type (e.g., normal, abnormal), and specific disease category.
Barrett's Grading	Grading of Barrett's esophagus severity: normal-barretts, barretts-short-segment, high-grade-dysplasia.
Ulcerative Colitis Grading	Inflammation severity scoring (MAYO-0, MAYO-1, MAYO-2).
Polyp Subtyping	Classification into the following subtypes: hyperplastic, adenomatous, serrated, or invasive-cancerous.
BBPS Scoring	Boston Bowel Preparation Scale scoring for colon cleanliness across three segments: BBPS-0-1 (well-prepared), BBSP-2-3 (not well-prepared), impacted-stool
Esophagus Disease Classification	Classification of esophageal diseases: esophagitis, esophageal-varices, esophageal-cancer, barretts, esophageal-candidiasis, esophageal-stricture, fistula-esophageal, esophageal-polyp, weiss-tear, diverticulum-esophageal, barretts-short-segment, achalasia, esophageal-caustic-injury.
Stomach Disease Classification	Classification of stomach diseases: portal-hypertensive-gastropathy, stomach-polyp, gastric-ulcer, erosion-gastric, gastric-varices, gastric-cancer, hyperemic-gastric, stomach-cancer, hemorrhagic-gastropathy, stenosis-stomach, gastric-lymphoma, hiatus-hernia, dieulafoy-lesion, gastric-caustic-injury
Small Intestine Disease Classification	Classification of small bowel conditions: erosion-SI, crohns, polyp-like, SI-ulcer, lymph-follicle, lymphangiectasia, angiectasia-SI, erythema, angiodysplasia-SI, blood-SI, stenosis-SI, submucosal-tumor, edema, intestinal-polyp, SI-cancer, hyperemic-SI, coeliac-disease, diverticulum-SI, duodenal-bulb-deformity
Colon Disease Classification	Classification of colonic diseases: colon-polyp, colorectal-cancer, ulcerative-colitis, colon-ulcer, angiodysplasia-colon, erosion-colon, diverticulosis, polyposis-syndrome, proctitis, lipoma, parasites, colon-stricture, fistula-colon, melanosis, hemorrhoids, ischemic-colitis, mucosal-inflammation, colon-diverticula, angiectasia-colon.
Stomach Anatomical Landmark Classification	Identification of anatomical landmarks in the stomach: non-landmark, pylorus, retroflex-stomach.

where w is the weight corresponding to the sample's class. For the GI-Tree classification task, the BCE loss was computed independently at each level of the hierarchy.

5. Results

We conducted a comprehensive evaluation of PanEndoFM, endoscopy-specific foundation models (EndoFM-LV, EndoSSL) and general-purpose vision models (ViT-B/16 and ResNet-50) on the PanEndoAtlas (Table 3). Performance was assessed using macro-AUC on the test set, which averages AUC across all classes to mitigate the effects of class imbalance and better

reflect model robustness in clinically heterogeneous settings.

We conducted a comprehensive evaluation of PanEndoFM, prior endoscopy-specific foundation models (EndoFM-LV, EndoSSL), and general-purpose vision models (ViT-B/16 and ResNet-50) on the PanEndoAtlas benchmark (Table 3). Performance was measured using macro-AUC, which averages AUC across all classes to mitigate class imbalance and provide a fairer measure of robustness in clinically heterogeneous settings. Across all 10 tasks, PanEndoFM ranked first on 6 tasks, second on 3 tasks, and third on 1 task, showing consistent superiority over both endoscopy-specific and general-purpose baselines.

PanEndoFM. PanEndoFM achieved the highest performance on the most comprehensive task, GI-Tree classification (111 classes, three-level hierarchy), with a macro-AUC of 0.648 (95% CI: 0.621–0.669), outperforming all other models. It also led performance on multiple upper GI tasks, including Barrett's grading (macro-AUC 0.775, 95% CI: 0.643–0.902) and esophageal disease classification (macro-AUC 0.772, 95% CI: 0.742–0.809). For stomach-related tasks, PanEndoFM achieved the best results in both disease classification (macro-AUC 0.539, 95% CI: 0.520–0.558) and anatomical landmark identification (macro-AUC 0.950, 95% CI: 0.943–0.956). It also topped ulcerative colitis grading with a macro-AUC of 0.571 (95% CI: 0.499–0.649). These results suggest that pretraining with broader anatomical coverage improves generalization across diverse disease types and organs.

Table 3: Macro-AUCs (95%CI) reported on test set for tasks in the PanEndoX Benchmark. Highest value is bolded, second highest is italicized.

Task Name	Classes	PanEndoFM	EndoFM-LV	EndoSSL	ViT-B/16	ResNet-50
GITree Classification	111 classes	**0.648 (0.621, 0.669)**	*0.570 (0.560, 0.578)*	0.551 (0.510, 0.599)	0.500 (0.491, 0.510)	0.515 (0.505, 0.526)
Barrett's Grading Classification	4 classes	**0.775 (0.643, 0.902)**	*0.733 (0.608, 0.850)*	0.581 (0.447, 0.719)	0.681 (0.559, 0.796)	0.707 (0.573, 0.843)
UCC Grading Classification	3 classes	**0.571 (0.499, 0.649)**	*0.565 (0.476, 0.650)*	0.507 (0.425, 0.583)	0.550 (0.470, 0.625)	0.525 (0.435, 0.614)
Polyp Subtyping Classification	4 classes	0.575 (0.562, 0.589)	0.500 (0.490, 0.511)	**0.704 (0.696, 0.712)**	0.542 (0.532, 0.552)	*0.643 (0.631, 0.655)*
BBPS Scoring Classification	3 classes	0.809 (0.773, 0.843)	**0.913 (0.885, 0.936)**	0.524 (0.479, 0.567)	*0.830 (0.800, 0.860)*	0.685 (0.648, 0.724)
Esophagus Disease Classification	13 classes	**0.772 (0.742, 0.809)**	0.590 (0.564, 0.612)	0.550 (0.519, 0.581)	*0.682 (0.645, 0.713)*	0.446 (0.418, 0.478)
Stomach Disease Classification	14 classes	**0.539 (0.520, 0.558)**	*0.539 (0.509, 0.582)*	0.521 (0.475, 0.554)	0.438 (0.410, 0.467)	0.498 (0.457, 0.536)
SI Disease Classification	20 classes	*0.594 (0.577, 0.612)*	0.481 (0.458, 0.507)	0.591 (0.573, 0.618)	**0.597 (0.577, 0.621)**	0.528 (0.509, 0.548)
Colon Disease Classification	30 classes	*0.552 (0.517, 0.586)*	**0.615 (0.583, 0.644)**	0.455 (0.430, 0.484)	0.494 (0.470, 0.519)	0.503 (0.467, 0.540)
Stomach Landmark Identification	3 classes	**0.950 (0.943, 0.956)**	0.869 (0.851, 0.886)	0.788 (0.749, 0.831)	*0.879 (0.866, 0.891)*	0.820 (0.800, 0.838)

EndoFM-LV. EndoFM-LV demonstrated strong performance on colon-focused tasks, outperforming all models on BBPS scoring (macro-AUC 0.913, 95% CI: 0.885–0.936) and colon disease classification (macro-AUC 0.615, 95% CI: 0.583–0.644). It also ranked second across five additional tasks. This advantage likely reflects its pretraining corpus, which was dominated

by colonoscopy videos, including HyperKvasir and SUN datasets. Because these datasets are also part of PanEndoAtlas, EndoFM-LV may benefit from pretraining-test overlap on colon-related tasks, inflating apparent generalization performance.

EndoSSL. EndoSSL performed best on polyp subtyping (macro-AUC 0.704, 95% CI: 0.696–0.712), consistent with its pretraining dataset being enriched for polyp-containing frames. However, it ranked outside the top two on most other tasks, particularly those involving non-colonic regions. This highlights the limitations of organ-specific pretraining strategies when models are applied to broader endoscopic practice.

General-purpose models. Interestingly, ViT-B/16 outperformed all models on small intestine disease classification (macro-AUC 0.597, 95% CI: 0.577–0.621) and achieved the second-highest performance on BBPS scoring, esophageal disease classification, and stomach landmark identification. We note that these tasks had least diversity in data sources.

Task-level difficulty. Macro-AUC values across multi-center tasks in PanEndoX were generally lower than those reported on single-center or task-specific datasets.[25] Easier tasks included anatomical landmark detection and bowel preparation scoring, where all models achieved relatively high performance. In contrast, grading-based tasks (e.g., ulcerative colitis grading, polyp subtyping) and multi-class disease classification tasks (e.g., GI-Tree) proved more difficult, with larger inter-model performance gaps. This spread highlights the value of PanEndoAtlas in exposing task difficulty and dataset diversity that are not captured by prior benchmarks.

Error analysis. Certain disease categories reflect intrinsic visual similarity rather than data scarcity and were often confused. For example, diverticulitis was frequently misclassified as diverticulosis, both involve diverticula; only inflamed ones count as diverticulitis. Similarly, ulcerative-colitis, ischemic-colitis, angiodysplastic-colitis, and infectious-colitis were often misclassified since all show inflamed, ulcerated mucosa, but with subtle cause-specific differences. Class-wise metrics for each task and model are provided in Supplementary Data.

Subgroup analysis. To assess generalization, we compared model performance on individual datasets against overall performance within each task. We observed substantial variance, with certain datasets consistently yielding higher metrics than others (Supplementary Data). The extent of this variance depended on the specific model–task combination, underscoring the strong influence of pretraining data on generalization performance. These findings demonstrate that reliance on single datasets can lead to overly optimistic performance estimates and reinforce the need for diverse, multi-center benchmarks.

Clinical interpretation. Although absolute improvements in macro-AUC may appear modest, they are clinically meaningful. Because macro-AUC weights rare and common diseases equally, even small gains can reduce missed diagnoses in less prevalent but high-risk conditions such as Barrett's esophagus or early gastric cancer. For tasks already approaching high accuracy (e.g., bowel preparation, anatomical landmark detection), incremental gains instead reflect greater robustness across institutions and devices.

6. Discussion

This work directly addresses the core limitations outlined in the introduction: the lack of large-scale, standardized, and clinically diverse benchmarks in endoscopy; the absence of clinically meaningful evaluation tasks; and the need to assess generalist versus specialist models in ways that reflect clinical reasoning.

First, PanEndoAtlas resolves the fragmentation of existing resources. Prior datasets have been narrowly focused, often on polyps, and inconsistent in their label vocabularies. By harmonizing 21 public datasets into a unified schema spanning 111 gastrointestinal abnormalities, PanEndoAtlas provides a standardized, hierarchical taxonomy that mirrors real diagnostic reasoning. This structure not only enables evaluation across rare and common conditions but also captures clinically relevant misclassifications that would be obscured in flat-label systems.

Second, PanEndoX introduces benchmarks grounded in clinical practice. The ten tasks span hierarchical classification, severity grading, disease subtyping, bowel preparation scoring, and anatomical landmark identification. Together, they probe the full spectrum of diagnostic reasoning in endoscopy—moving beyond single-center, polyp-focused tasks toward clinically meaningful evaluation. Our results demonstrate that task difficulty varies systematically, with multi-class grading and disease classification posing greater challenges than landmark detection or bowel preparation scoring. This reinforces the importance of diverse benchmarks for exposing gaps in model generalization.

Third, PanEndoFM demonstrates the value of broad, multi-organ pretraining for endoscopic AI. While prior models like EndoSSL and EndoFM-LV show strong performance on niche domains (polyp subtyping and colon disease tasks, respectively), PanEndoFM outperformed both endoscopy-specific and general-purpose models on six of ten tasks. Therefore, PanEndoSuite provides a rigorous foundation for evaluating generalization and robustness, thereby bridging AI development with real-world practice.

Limitations and future directions. We note that some prior endoscopy foundation models incorporated public datasets included in PanEndoAtlas for pretraining, even if labels were not used, partially limiting the independence of certain comparisons. Future work should establish clear dataset disjointness to enable standardized evaluation. We further note that a few of the images in the public datasets are low quality. We provide auto-generated quality annotations for each image in PanEndoAtlas.

7. Data and Code Availability

The dataset, code and supplementary data is publicly available is publicly available at: https://github.com/rajpurkarlab/panendosuite.

8. Acknowledgements

This work was supported by Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence.

References

1. R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams and F. Mahmood, Towards a general-purpose foundation model for computational pathology, *Nature Medicine* **30**, p. 850–862 (March 2024).
2. M. Y. Lu, B. Chen, A. Zhang, D. F. K. Williamson, R. J. Chen, T. Ding, L. P. Le, Y.-S. Chuang and F. Mahmood, Visual language pretrained multiple instance zero-shot transfer for histopathology images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.
3. M. Paschali, Z. Chen, L. Blankemeier, M. Varma, A. Youssef, C. Bluethgen, C. Langlotz, S. Gatidis and A. Chaudhari, Foundation models in radiology: What, how, why, and why not, *Radiology* **314** (February 2025).
4. P. Neidlinger, O. S. M. E. Nahhas, H. S. Muti, T. Lenz, M. Hoffmeister, H. Brenner, M. van Treeck, R. Langer, B. Dislich, H. M. Behrens, C. Röcken, S. Foersch, D. Truhn, A. Marra, O. L. Saldanha and J. N. Kather, Benchmarking foundation models as feature extractors for weakly-supervised computational pathology (2024).
5. X. Zhang, H.-Y. Zhou, X. Yang, O. Banerjee, J. N. Acosta, J. Miller, O. Huang and P. Rajpurkar, Rexrank: A public leaderboard for ai-powered radiology report generation (2024).
6. D. A. Fisher, A. K. Shergill, D. S. Early, R. D. Acosta, V. Chandrasekhara, K. V. Chathadi, G. A. Decker, J. A. Evans, R. D. Fanelli, K. Q. Foley, L. Fonkalsrud, J. H. Hwang, T. Jue, M. A. Khashab, J. R. Lightdale, V. R. Muthusamy, S. F. Pasha, J. R. Saltzman, R. Sharaf and B. D. Cash, Role of endoscopy in the staging and management of colorectal cancer, *Gastrointestinal Endoscopy* **78**, p. 8–12 (July 2013).
7. L. Zhu, J. Qin, J. Wang, T. Guo, Z. Wang and J. Yang, Early gastric cancer: Current advances of endoscopic diagnosis and treatment, *Gastroenterology Research and Practice* **2016**, p. 1–7 (2016).
8. N. H. Kim, Y. S. Jung, W. S. Jeong, H.-J. Yang, S.-K. Park, K. Choi and D. I. Park, Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies, *Intestinal Research* **15**, p. 411 (2017).
9. A. Turshudzhyan, H. Rezaizadeh and M. Tadros, Lessons learned: Preventable misses and near-misses of endoscopic procedures, *World Journal of Gastrointestinal Endoscopy* **14**, p. 302–310 (May 2022).
10. R. Hirsch, M. Caron, R. Cohen, A. Livne, R. Shapiro, T. Golany, R. Goldenberg, D. Freedman and E. Rivlin, Self-supervised learning for endoscopic video analysis (2023).
11. Z. Wang, C. Liu, S. Zhang and Q. Dou, Foundation model for endoscopy video analysis via large-scale self-supervised pre-train (2023).
12. Z. Wang, C. Liu, L. Zhu, T. Wang, S. Zhang and Q. Dou, Improving foundation model for endoscopy video analysis via representation learning on long sequences, *IEEE Journal of Biomedical and Health Informatics* , 1 (2025).
13. Q. Tian, H. Liao, X. Huang, B. Yang, D. Lei, S. Ourselin and H. Liu, Endomamba: An efficient foundation model for endoscopic videos (2025).
14. P. Dermyer, A. Kalra and M. Schwartz, Endodino: A foundation model for gi endoscopy (2025).
15. S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen, A. Petlund, P. Halvorsen, J. Rittscher, T. de Lange and J. E. East, A multi-centre polyp detection and segmentation dataset for generalisability assessment, *Scientific Data* **10** (February 2023).
16. D. Jha, N. K. Tomar, V. Sharma, Q.-H. Trinh, K. Biswas, H. Pan, R. K. Jha, G. Durak, A. Hann, J. Varkey, H. V. Dao, L. Van Dao, B. P. Nguyen, N. Papachrysos, B. Rieders, P. T. Schmidt,

E. Geissler, T. Berzin, P. Halvorsen, M. A. Riegler, T. de Lange and U. Bagci, Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy (2024).
17. M. Forootan, M. Rajabnia, A. R. Mafi, H. A. Tehrani, E. Ghadirzadeh, M. Setayeshfar, Z. Ghaffari, M. Tashakoripour, M. R. Zali and H. Bolhasani, Ercpmp: An endoscopic image and video dataset for colorectal polyps morphology and pathology (2023).
18. worldendo.org
`https://www.worldendo.org/assets-craft/pdf/resources/mst-3-0.pdf`, [Accessed 02-08-2025].
19. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin and P. Bojanowski, Dinov2: Learning robust visual features without supervision (2023).
20. J. Silva, A. Histace, O. Romain, X. Dray and B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, *International Journal of Computer Assisted Radiology and Surgery* **9**, p. 283–293 (September 2013).
21. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez and F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, *Computerized Medical Imaging and Graphics* **43**, p. 99–111 (July 2015).
22. J. Cychnerski, T. Dziubich and A. Brzeski, Ers: a novel comprehensive endoscopy image dataset for machine learning, compliant with the mst 3.0 specification (2022).
23. S. Zhu, J. Gao, L. Liu, M. Yin, J. Lin, C. Xu, C. Xu and J. Zhu, Public imaging datasets of gastrointestinal endoscopy for artificial intelligence: a review, *Journal of Digital Imaging* **36**, p. 2578–2601 (September 2023).
24. K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler and P. Halvorsen, Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17 (ACM, June 2017).
25. H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen and T. de Lange, Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, *Scientific Data* **7** (August 2020).
26. Y. Ma, X. Chen, K. Cheng, Y. Li and B. Sun, *LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps*, in *Lecture Notes in Computer Science*, (Springer International Publishing, 2021), p. 387–396.
27. K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, M. Riegler and P. Halvorsen, Nerthus: A bowel preparation quality video dataset, in *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17 (ACM, June 2017).
28. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen and H. D. Johansen, Kvasir-seg: A segmented polyp dataset (2019).
29. S. Ali, B. Braden, D. Lamarque, S. Realdon, A. Bailey, R. Cannizzaro, N. Ghatwary, J. Rittscher, C. Daul and J. East, Endoscopy disease detection and segmentation (edd2020) (2020).
30. W. Wang, J. Tian, C. Zhang, Y. Luo, X. Wang and J. Li, An improved deep learning approach and its applications on colonic polyp images detection, *BMC Medical Imaging* **20** (July 2020).
31. P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler and P. Halvorsen, Kvasir-capsule, a video capsule endoscopy dataset, *Scientific Data* **8** (May

2021).
32. D. Jha, N. K. Tomar, S. Ali, M. A. Riegler, H. D. Johansen, D. Johansen, T. de Lange and P. Halvorsen, Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy (2021).
33. D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen and M. A. Riegler, A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation, *IEEE Journal of Biomedical and Health Informatics* **25**, p. 2029–2040 (June 2021).
34. K. Li, M. I. Fathan, K. Patel, T. Zhang, C. Zhong, A. Bansal, A. Rastogi, J. S. Wang and G. Wang, Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations, *PLOS ONE* **16**, p. e0255809 (August 2021).
35. M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda and K. Mori, Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video), *Gastrointestinal Endoscopy* **93**, 960 (April 2021).
36. A. de Maissin, R. Vallée, M. Flamant, M. Fondain-Bossiere, C. L. Berre, A. Coutrot, N. Normand, H. Mouchère, S. Coudol, C. Trang and A. Bourreille, Multi-expert annotation of crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network, *Endoscopy International Open* **09**, p. E1136–E1144 (June 2021).
37. G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen and L. Van Gool, Video polyp segmentation: A deep learning perspective, *Machine Intelligence Research* **19**, p. 531–549 (November 2022).
38. D. Jha, V. Sharma, N. Dasu, N. K. Tomar, S. Hicks, M. K. Bhuyan, P. K. Das, M. A. Riegler, P. Halvorsen, U. Bagci and T. de Lange, Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection (2023).
39. D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, Q. Duan, J. Zhao, K. Li, Y. Qiao and S. Zhang, A real-world dataset and benchmark for foundation model adaptation in medical image classification, *Scientific Data* **10** (September 2023).
40. A. Yokote, J. Umeno, K. Kawasaki, S. Fujioka, Y. Fuyuno, Y. Matsuno, Y. Yoshida, N. Imazu, S. Miyazono, T. Moriyama, T. Kitazono and T. Torisu, Small bowel capsule endoscopy examination and open access database with artificial intelligence: The see-artificial intelligence project, *DEN Open* **4** (June 2023).
41. Dekker, ClinicalTrials.gov — clinicaltrials.gov `https://clinicaltrials.gov/study/NCT03822390#more-information`, (2020), [Accessed 06-05-2024].
42. A. Gu and T. Dao, Mamba: Linear-time sequence modeling with selective state spaces (2023).
43. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need (2017).
44. P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux and A. Bartoli, Computer-aided classification of gastrointestinal lesions in regular colonoscopy, *IEEE Transactions on Medical Imaging* **35**, p. 2051–2063 (September 2016).
45. A. Charoen, A. Guo, P. Fangsaard, S. Taweechainaruemitr, N. Wiwatwattana, T. Charoenpong and H. G. Rich, Rhode island gastroenterology video capsule endoscopy data set, *Scientific Data* **9** (October 2022).