Computer-Assisted Intervention in Capsule Endoscopy: A Real-Time Edge-AI Auditing System

Diego Bravo[1 [0000−0003−1957−1615]], Josué Ruano-Balseca[2,1 [0000−0003−4801−8229]], Fabio A. González[3 [0000−0001−9009−7288]], and Eduardo Romero[1]

[1] CIM@LAB, Universidad Nacional de Colombia, Bogotá, Colombia
[2] imec-Vision Lab, University of Antwerp, Antwerp, Belgium
[3] MindLab, Universidad Nacional de Colombia, Bogotá, Colombia
edromero@unal.edu.co

Abstract. Video capsule endoscopy (VCE) is the preferred first-line modality for visualizing the small bowel mucosa and evaluating obscure gastrointestinal bleeding. However, each examination generates over 50 thousand frames, making manual review time-consuming and prone to missed findings and inter-observer variability. Most AI systems rely on single-frame analysis, overlooking temporal dynamics and limiting their ability to distinguish anatomical transitions from transient artifacts. We propose a real-time auditing framework that jointly identifies the anatomical region along the gastrointestinal tract and detects anomalies through intrinsic temporal modeling. Consecutive frames are encoded using an endoscopy-pretrained CNN and processed by a lightweight Vision Transformer operating in a sequence-to-sequence manner, enabling frame-level predictions without post-processing. Evaluated on the multi-center Galar dataset (∼3.5 million frames), the proposed system achieves a macro F1-score of 90% for anatomical classification and 67% macro recall for binary anomaly detection within a unified multi-task learning framework, outperforming prior frame-wise approaches. Deployment on a low-power edge device confirms real-time performance (30.5 ms per frame, ∼33 FPS), enabling continuous monitoring of incomplete procedures, capsule retention, and persistent abnormalities. Our code is available at https://github.com/DiegoBravoH/CapsuleAudit.

Keywords: Video capsule endoscopy · Temporal modeling · Real-time.

1 Introduction

Video capsule endoscopy (VCE) is a minimally invasive diagnostic modality that enables visualization of the gastrointestinal (GI) tract, particularly the small intestine, which is difficult to assess by standard gastroscopy and colonoscopy [1]. Patients swallow a pill-sized wireless camera that passively traverses the GI tract, capturing images for subsequent review. VCE is currently recommended as the first-line examination for obscure gastrointestinal bleeding (OGIB) and

other small bowel disorders [2]. Each examination typically generates 50,000–80,000 images, requiring 40–120 minutes of focused interpretation [3]. Diagnostic accuracy depends heavily on sustained expert attention, and prolonged review may increase variability and the risk of missed findings, with reported detection rates of only ∼46% and moderate interobserver agreement ($\kappa \approx 0.44$) [4]. Despite technological advances in VCE [5], these limitations persist [6], underscoring the need for real-time computational support through AI-assisted procedural auditing.

Artificial intelligence (AI) is increasingly being integrated into VCE for assisted lesion detection and anatomical classification, with the potential to improve diagnostic consistency and reduce interobserver variability [7,8]. However, much of the reported performance is derived from retrospective evaluations or carefully curated datasets rather than full-length examinations under routine clinical conditions, limiting the reliability and generalizability of current systems [9]. Methodologically, most current approaches rely on single-frame classification using Convolutional Neural Networks (CNNs) to detect anatomical regions or pathological abnormalities [10,11,12]. Temporal consistency is typically enforced through post-processing techniques, such as Hidden Markov Models (HMMs) to constrain anatomical transitions [13,14]. Although performance improves, frame-level predictions remain unstable without post-processing, reflecting limited intrinsic temporal modeling. Furthermore, random sampling and class balancing strategies disrupt the natural temporal distribution of VCE data. Consequently, temporal coherence is imposed heuristically rather than learned directly from sequential data, limiting robust sequence-level modeling and the ability to capture long-range dependencies spanning anatomical transitions and subtle pathological evolution throughout consecutive frames in real-world VCE examinations.

This work presents a real-time computer-assisted intervention system for video capsule endoscopy (VCE) that jointly performs anatomical localization and anomaly detection. Operating at 30.5 ms/frame on an NVIDIA Jetson Nano, the system enables continuous auditing of capsule progression, examination completeness, potential capsule retention, and abnormal findings. By combining multi-task learning and temporal modeling, the proposed approach outperforms frame-wise methods without heuristic post-processing. Code, trained models, and a real-time prototype are publicly available to support reproducibility.

2 Method

Multi-Frame Feature Extraction. Given a video dataset composed of clips (temporal windows), a batch of input segments is denoted as: $X \in \mathbb{R}^{B\times 3\times F\times H\times W}$ being B the batch size, F the number of frames (tokens) per segment, and $H \times W$ the spatial resolution. Each frame is independently processed by a single pre-trained ResNet-50 backbone [15]. Let $x_{i,t} \in \mathbb{R}^{3\times H\times W}$ denote the *t-th* frame of the *i-th* segment. The CNN encoder maps each frame into a D-dimensional embedding (with $D = 2048$): $z_{i,t} = f_{\text{ResNet50}}(x_{i,t}) \in \mathbb{R}^D$. The frame-level embed-

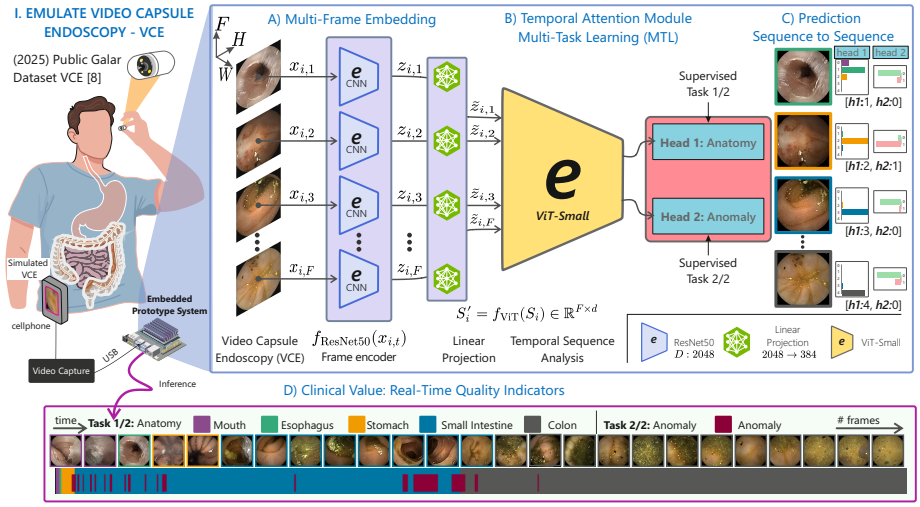

Fig. 1. Overview of the proposed multi-frame VCE auditing framework. Consecutive frames are independently encoded by a single pretrained CNN with shared weights, applied sequentially to each frame, into D-dimensional embeddings, projected and processed by a temporal Vision Transformer in a sequence-to-sequence manner. The model produces frame-wise predictions for anatomical localization and binary anomaly detection, which are aggregated to generate real-time clinical quality indicators.

dings are stacked along the temporal dimension to form the ordered sequence: $Z_i = [z_{i,1}, z_{i,2}, \ldots, z_{i,F}] \in \mathbb{R}^{F \times D}$, which serves as input to the projection layer (see Fig. 1A).

Sequence-to-Sequence Temporal Modeling. The CNN embeddings are projected through a trainable linear layer to match the transformer embedding dimension: $\tilde{z}_{i,t} = W z_{i,t} \in \mathbb{R}^d, \quad d = 384$. The projected sequence is therefore $S_i = [\tilde{z}_{i,1}, \ldots, \tilde{z}_{i,F}] \in \mathbb{R}^{F \times d}$. To capture temporal dependencies, we employ a Vision Transformer (ViT-Small [16]) operating over this sequence. Unlike standard ViT architectures that tokenize spatial patches, our model treats each projected frame embedding as a temporal token. The transformer produces a refined representation: $S_i' = f_{\text{ViT}}(S_i) \in \mathbb{R}^{F \times d}$. Temporal interactions are modeled using Multi-Head Self-Attention (MSA [17]): $\text{Attention}(Q, K, V) = \text{softmax}(QK^\top / \sqrt{d_k})V$, where Q, K, V are learned linear projections. We perform token-level prediction without using a CLS token. Each temporal token outputs a frame-level prediction rather than a single label per sequence, preserving temporal resolution and enabling real-time inference (Fig. 1B–C).

Task Definition and Report Quality Indicators. To enable automated auditing of VCE examinations, we address two tasks. Anatomical localization supports monitoring of capsule progression, detection of retrograde events, and

early identification of potential capsule retention, while anomaly detection identifies clinically relevant findings during the examination. **1) Anatomical section classification:** For each frame t in segment i, the model predicts $y_{i,t}^{\text{anat}} \in \{1, \ldots, C\}$, where $C = 5$ (mouth, esophagus, stomach, small intestine, colon). **2) Binary Anomaly Detection:** The model simultaneously predicts $y_{i,t}^{\text{anom}} \in \{0, 1\}$ (see Fig. 2). Using fixed-length sliding windows (stride 1), each frame appears in multiple windows. Let $\mathbf{p}_{i,t}^{k}$ denote the predicted probability vector for frame t in window k. Final frame-level probabilities are obtained by averaging the predicted probability vectors across all windows (soft voting over probabilities). Uncertainty is estimated via Shannon entropy computed over the averaged probabilities: $H_{i,t} = -\sum_c \bar{p}_c \log(\bar{p}_c + \epsilon)$, where ϵ ensures numerical stability. The entropy is normalized by division by $\log C$ to scale it to $[0, 1]$. Higher entropy indicates temporal disagreement and supports frame-level quality assessment.

3 Experiments and Results

3.1 Experimental Setup

Datasets. Experiments were conducted on the recently released Galar video capsule endoscopy dataset [18] (2025). Galar comprises 80 fully annotated VCE studies with 3,513,539 frame-wise labeled images collected from two centers and multiple capsule systems (Olympus$^{\text{TM}}$ and PillCam$^{\text{TM}}$). In this work, we directly use the provided frame-level annotations without modification and adopt the official split of 60 studies for training/validation and 20 for testing, ensuring patient-level separation. To evaluate cross-dataset robustness, we additionally tested the trained anatomical model on the Rhode Island GI VCE dataset [19]. The official test split was used for inference without any fine-tuning, employing the anatomical model trained on Galar (Cases 1–60). As this study uses open-access human subject data released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, no additional ethical approval was required.

Implementation Details. Both the CNN and transformer backbones are initialized from GastroNet-5M foundation models [15], pretrained using self-supervised DINO learning on >5 million gastrointestinal endoscopy images collected from eight clinical centers [20]. The temporal encoder employs a ViT-Small backbone ($\sim 21.7M$ parameters) adapted to process variable length temporal sequences ($F \in [5, 195]$ frames). Each frame (resized to 224×224) is encoded by a ResNet50 backbone ($\sim 23.5M$ parameters), producing a 2048-dimensional embedding. These embeddings are projected via a trainable linear layer (∼0.79M parameters) into a 384-dimensional token representation before being fed into the transformer encoder. The model follows a sequence-to-sequence formulation, generating frame-wise predictions for both anatomical classification (5 classes: mouth, esophagus, stomach, small intestine, colon) and binary anomaly detection (see Fig. 1-C). The classification heads consist of two parallel MLP layers operating on the refined temporal representations. Training is performed in two

phases. During a warm-up stage, only the projection and classification heads are trained for 10 epochs using a fixed learning rate of 1×10^{-3} to stabilize convergence. In the fine-tuning stage, all layers are unfrozen and optimized for up to 30 additional epochs, with early stopping based on validation macro F1-score (patience = 5 epochs). To balance the Multi-Task Learning (MTL), we employ homoscedastic uncertainty weighting [21], which learns task-specific log-variance parameters to automatically scale each loss: $\mathcal{L} = e^{-s_a}\mathcal{L}_{\text{anom}} + s_a + e^{-s_b}\mathcal{L}_{\text{anat}} + s_b$, where s_a and s_b are learnable parameters that adaptively adjust the relative contribution of each task, avoiding manual loss weighting. The model is implemented in PyTorch. Optimization is performed using Adam with a weighted Focal Loss ($\gamma = 2.0$) to mitigate class imbalance and emphasize hard-to-classify samples. The learning rate is reduced by a factor of $\gamma = 0.1$ every 3 epochs. Training was conducted on a single NVIDIA RTX 4500 GPU. The temporal transformer encoder operates on fixed temporal windows and is designed for causal inference without future frames.

Ablation Study. This study was conducted to evaluate the contribution of the temporal modeling component to both anatomical classification and binary anomaly detection. Specifically, we compared the full sequence-to-sequence model across different temporal window sizes against a frame-wise baseline (see Tables 1 and 2). We further evaluated temporal post-processing (smoothing) to assess its effect on anatomical prediction stability (see Table 1, HMM column). Importantly, post-processing was evaluated but is not considered part of the core framework, as the primary objective of this study is to preserve real-time streaming inference on an edge AI device.

3.2 Model Validation

Task 1/2 Anatomical. Table 1 summarizes the performance for anatomical classification under single-frame and multi-frame settings. In the single-frame analysis, linear probing using endoscopy-pretrained embeddings (GastroNet [20,15]) significantly outperformed ImageNet-pretrained features (macro F1: 77% vs. 51%), underscoring the importance of domain-specific pretraining. Temporal modeling significantly improved performance across all anatomical classes under both Single-Task Learning (STL) and Multi-Task Learning (MTL), compared to state-of-the-art frame-wise approaches [12,14]. Performance gains were most pronounced in mouth, followed by esophagus, stomach, colon, and small intestine (S. Bowel). Overall, these results demonstrate that intrinsic temporal modeling improves anatomical consistency over frame-wise approaches. Werner et al. [14] serves as our primary baseline, reporting single-frame MTL results with and without HMM post-processing; we compare our model against their best reported performance. In contrast, we retain all samples to preserve the intrinsic temporal structure and natural class distribution of VCE data. Clinical VCE data are inherently imbalanced, altering this distribution may limit the representativeness of real-world performance. In the STL setting, an ablation study compared CNN+GRU (Gated Recurrent Unit) and CNN+ViT (with

Table 1. Performance comparison between single-frame and multi-frame approaches for anatomical classification. Models were trained/validated in 60 studies (∼3.2M frames) and evaluated in 20 independent test studies (∼234K frames). 95% confidence intervals were calculated using stratified bootstrapping with 1,000 resamples by case. Results are reported as class-wise F1-scores and macro F1. HMM denotes Hidden Markov Model post-processing and the column reports the resulting macro F1-score. STL: Single-Task Learning (task: anatomy) and MTL: Multi-Task Learning (tasks: anatomy, anomaly). The shaded row (MTL-75) denotes the model selected for edge deployment.

Ablation Study F1-score(%) — Anatomical Task — Single-frame							
Token=1	Mouth	Esophag.	Stomach	S. Bowel	Colon	Macro F1	HMM
ImageNet	10.1±5.0	26.7±15.6	61.2±11.5	90.0±3.6	68.5±12.1	51.3±6.0	88.7±6.5
GastroNet	60.9±11.3	69.0±18.4	82.4±7.5	94.0±3.5	81.0±10.4	77.5±5.8	93.3±4.4
Tokens≥5 — Endoscopy-pretrained CNN+ViT (Multi-frame)							
STL 5	80.7±8.7	87.2±9.9	89.6±4.6	95.3±3.4	85.6±11.1	87.7±5.0	93.7±4.9
STL 31	**88.8±6.4**	**91.1±7.1**	**93.7±3.5**	94.5±3.6	81.8±12.0	90.0±4.6	93.9±3.9
STL 75	88.4±6.5	89.8±6.6	93.3±4.3	95.4±3.7	84.6±11.7	90.3±5.0	**93.9±3.9**
STL 195	87.4±7.2	89.3±7.8	93.3±3.7	**96.0±3.6**	**87.3±12.0**	**90.7±4.7**	93.2±4.2
MTL 5	58.7±20.2	72.8±17.5	85.6±6.5	94.5±3.5	82.8±10.6	78.9±6.5	92.4±5.1
MTL 31	70.0±19.1	89.7±7.7	93.1±3.5	94.4±3.8	81.7±11.8	85.8±6.5	93.8±4.0
MTL 75	87.0±7.6	89.5±7.7	92.1±4.7	95.9±3.5	86.9±11.2	90.3±5.3	93.0±4.2
MTL 195	22.9±12.1	67.4±18.2	92.5±4.4	95.4±3.5	85.5±11.2	72.8±6.4	92.4±4.1

Comparison with state-of-the-art single-frame approaches (same dataset)							
Token=1	Mouth	Esophag.	Stomach	S. Bowel	Colon	Macro F1	HMM
[12] 2025	42.0	65.0	78.0	93.0	75.0	71.0	-
[14] 2025	-	-	-	-	-	65.1	92.4

randomly initialized ViT) models across different temporal window sizes. Macro F1-scores (%) for 5, 31, and 75 tokens were 87 vs. 89, 89 vs. 89, and 85 vs. 89, respectively.

Similar experiments under the MTL configuration yielded macro F1-scores (%) of 73 vs. 74 (5 tokens), 89 vs. 67 (31 tokens), and 88 vs. 81 (75 tokens), showing that endoscopy-pretrained ViT initialization improved performance overall (see Table 1). Given the clinical objective of minimizing missed abnormalities, model selection prioritized sensitivity in the anomaly detection task while maintaining strong anatomical classification performance within a single multi-task framework. Based on these criteria, MTL-75 was selected as the final deployment model across all evaluated window sizes and learning settings. Although STL-195 achieved the highest anatomical macro F1-score (90.7%), MTL-75 achieved comparable anatomical classification performance (90.3%) while providing the highest anomaly macro recall (66.7%, see Table 2).

Post-Processing – Hidden Markov Model. The HMM was replicated as reported in the baseline study [14], confirming performance gains in all cases, particularly for single-frame models (see Table 1, HMM column). However, it assumes a strictly forward transition sequence (Mouth → Esophagus → Stomach → Small Intestine → Colon) with a fixed initial state at the mouth. Although

anatomically plausible, this constraint biases early predictions and may distort minority classes. Moreover, as a Viterbi-based post-processing step requiring full-sequence access, it is not suitable for real-time deployment. Clinically, enforcing a unidirectional structure may obscure retrograde events (e.g., gastro-esophageal or duodeno-gastric reflux, linked to Barrett's metaplasia and cancer [22]; see reflux case 28, Galar dataset), potentially masking relevant pathological findings.

Robustness. The **MTL-75** model was cross-evaluated with the Rhode Island GI VCE test set (∼1M frames) [10] without fine-tuning or post-processing. It achieved F1-scores (%) of 68 (esophagus), 91 (stomach), 94 (small intestine), and 62 (colon). As the *mouth* class is not annotated in this dataset, all predictions assigned to this category (n=347) were treated as false positives. Performance degradation in the esophagus and colon likely reflects domain shift between datasets.

Task 2/2 Anomaly. Table 2 reports binary anomaly detection performance following the state-of-the-art evaluation protocol of [23,14], using the same dataset, grouping strategy (**Normal:** Clean Mucosa; **Anomaly:** Polyp, Blood, Active Bleeding, Angiectasia, Erosion, Erythema, Ulcer), and test set (patient IDs 61–80). In clinical practice, sensitivity is critical, as missing an anomaly is more detrimental than generating a false positive alert. Under this criterion, the best-performing model was **MTL-75**, with endoscopy-pretrained initialization of both CNN and ViT backbones, achieving a macro recall of 67% and weighted recall/precision/F1 of 77%/81%/79%, respectively. This model was therefore selected for deployment on the edge AI device. In comparison, the same CNN+ViT architecture evaluated under STL/MTL settings with alternative temporal window sizes (5, 31, and 195 tokens) achieved macro F1-scores (%) of 52/59 (5 tokens), 55/62 (31 tokens), and 59/55 (195 tokens), respectively. A major challenge is frame-level label inconsistency (e.g., anomaly–normal–anomaly sequences, see Fig. 2 anomaly ground truth and entropy). Anomalies frequently occupy only a small portion of the frame, while visual artifacts such as bubbles or luminal fluids can obscure mucosal details, making consistent annotation difficult.

Table 2. Performance comparison of single-frame and multi-frame approaches for binary anomaly detection using deep learning architectures. Same data split and 95% CI estimation as in Table 1. The shaded row (MTL-75) corresponds to the final deployment model used in the edge AI prototype. **Note:** Anomaly category was extended to include inflammatory bowel disease, foreign body, hematin, cancer, and lymphangiectasis, as these findings are clinically considered abnormal.

Task 2/2 Anomaly — Single-frame and Multi-frame (STL, MTL)						
Method	Micro F1	Macro F1	Macro Prec.	Macro Rec.	Model	Embedded
[14] 2025	**87.7**	54.4	54.1	54.7	ResNet50	No
[23] 2025	87.3	37.0	26.6	60.9	MobilNet-S	No
GastroNet	79.5±6.9	61.6±8.4	61.7±8.4	62.4±9.1	ResNet50	Yes
STL 75	85.7±9.3	59.7±11.1	**76.8±15.7**	58.2±7.8	CNN+ViT	Yes
MTL 75	77.0±6.2	**62.8±10.3**	61.9±9.4	**66.7±13.8**	CNN+ViT	Yes

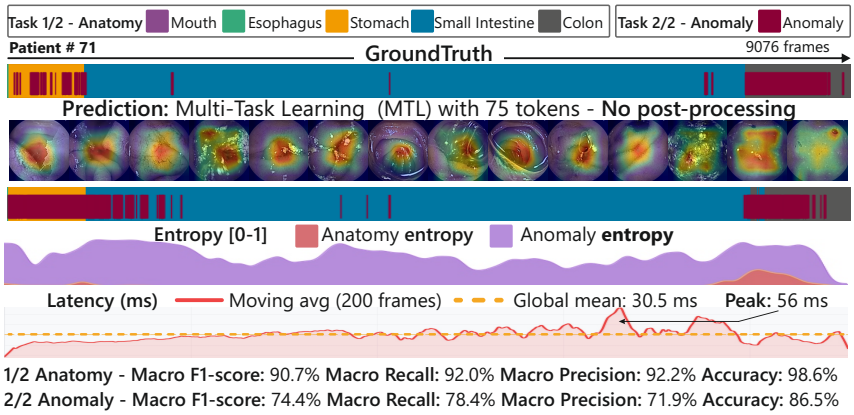

Fig. 2. Quality indicator report for Patient #71. Ground-truth and MTL-75 frame-level predictions (no post-processing) for anatomy and anomaly tasks are shown, along with entropy-based uncertainty and edge inference latency (Jetson Nano).

3.3 Real-Time Deployment on an Edge AI Device: Performance Analysis and Practical Considerations

To evaluate feasibility beyond retrospective offline benchmarking, the proposed CNN+ViT multi-task framework was deployed on a low-power NVIDIA Jetson Nano edge device (100×80 mm, 5–10 W) to assess real-time performance. The system integrates video acquisition (HDTV capture card with smartphone-based signal emulation, Fig. 1-I), feature extraction (ResNet50 backbone), temporal modeling (75-frame window), and multi-task prediction (anatomy and anomaly) within a fully parallelized pipeline. The trained PyTorch model (.pth and .ckpt) was converted to Open Neural Network eXchange (ONNX) and subsequently optimized using TensorRT. For real-time operation, temporal windows contain only past and present frames, ensuring causal online inference without future-frame access. Each frame is encoded once by the ResNet50 backbone and reused across overlapping windows (stride = 1), while the ViT-Small temporal encoder produces frame-level (token) prediction within the window. Predictions from overlapping windows are aggregated through soft voting with entropy-based uncertainty estimation (see Section 2). As illustrated in Figure 2 (Patient #71), stable frame-level predictions are maintained throughout the examination (9,076 frames) with an average latency of 30.5 ms per frame (≈33 FPS), exceeding the typical 2–6 FPS acquisition rate of standard small-bowel VCE systems [24]. Importantly, deployment does not rely on post-processing (e.g., HMM smoothing), demonstrating that temporal consistency emerges directly from learned representations. These results demonstrate operational feasibility for intra-procedural, edge-based deployment, supporting on-device monitoring of examination completeness and potential capsule retention, which occur in 0.7–1.4% of examinations and contribute to incomplete visualization (∼12%) [25]. Such real-time monitoring is particularly relevant in high-risk populations, including pediatric

and inflammatory bowel disease patients [7]. Further clinical and system-level validation is required prior to integration into routine practice. Code and trained models are publicly available.

4 Conclusions

We present a real-time, edge-deployable system for video capsule endoscopy that intrinsically models temporal dependencies to jointly perform anatomical localization and binary anomaly detection at the device's native acquisition rate, without heuristic post-processing. Although evaluated on a multi-center, multi-device dataset (Galar), the cohort was predominantly pathological, which may limit generalization to fully normal examinations. Further validation on larger and more diverse adult and pediatric cohorts is warranted. Future work will focus on clinical validation studies and the development of a standardized validation and alerting protocol for real-time deployment, including expert-defined thresholds for incomplete examinations, prolonged segmental transit, capsule retention, and persistent anomalies. In parallel, advances in spatiotemporal modeling, interpretability, capsule-specific foundation models for VCE, and alternative sequence architectures (e.g., state-space models such as Mamba) will be explored to improve robustness and energy-efficient deployment.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Fisher, L., Krinsky, M.L., Anderson, M.A., Appalaneni, V., Banerjee, S., Ben-Menachem, T., Cash, B.D., Decker, G.A., Fanelli, R.D., Friis, C., et al.: The role of endoscopy in the management of obscure gi bleeding. Gastrointestinal endoscopy **72**(3), 471–479 (2010)
2. Raju, G.S., Gerson, L., Das, A., Lewis, B.: American gastroenterological association (aga) institute medical position statement on obscure gastrointestinal bleeding. Gastroenterology **133**(5), 1694–1696 (2007)
3. Mishkin, D.S., Chuttani, R., Croffie, J., DiSario, J., Liu, J., Shah, R., Somogyi, L., Tierney, W., Song, L.M.W.K., Petersen, B.T.: Asge technology status evaluation report: wireless capsule endoscopy. Gastrointestinal endoscopy **63**(4), 539–545 (2006)
4. Rondonotti, E., Soncini, M., Girelli, C.M., Russo, A., Ballardini, G., Bianchi, G., Cantù, P., Centenara, L., Cesari, P., Cortelezzi, C.C., et al.: Can we improve the detection rate and interobserver agreement in capsule endoscopy? Digestive and Liver Disease **44**(12), 1006–1011 (2012)
5. Rahman, M., Akerman, S., DeVito, B., Miller, L., Akerman, M., Sultan, K.: Comparison of the diagnostic yield and outcomes between standard 8 h capsule endoscopy and the new 12 h capsule endoscopy for investigating small bowel pathology. World Journal of Gastroenterology: WJG **21**(18), 5542 (2015)
6. Wang, Y.C., Pan, J., Liu, Y.W., Sun, F.Y., Qian, Y.Y., Jiang, X., Zou, W.B., Xia, J., Jiang, B., Ru, N., et al.: Adverse events of video capsule endoscopy over the past two decades: a systematic review and proportion meta-analysis. BMC gastroenterology **20**(1), 364 (2020)

7. Rojas, I., Barth, B.A., Stewart, J.W.: Advances in pediatric video capsule endoscopy: current applications and future directions. Frontiers in Pediatrics **13**, 1738998 (2026)
8. Messmann, H., Bisschops, R., Antonelli, G., Libânio, D., Sinonquel, P., Abdelrahim, M., Ahmad, O.F., Areia, M., Bergman, J.J., Bhandari, P., et al.: Expected value of artificial intelligence in gastrointestinal endoscopy: European society of gastrointestinal endoscopy (esge) position statement. Endoscopy **54**(12), 1211–1231 (2022)
9. Piccirelli, S., Salvi, D., Pugliano, C.L., Tettoni, E., Facciorusso, A., Rondonotti, E., Mussetto, A., Fuccio, L., Cesaro, P., Spada, C.: Unmet needs of artificial intelligence in small bowel capsule endoscopy. Diagnostics **15**(9), 1092 (2025)
10. Charoen, A., Guo, A., Fangsaard, P., Taweechainaruemitr, S., Wiwatwattana, N., Charoenpong, T., Rich, H.G.: Rhode island gastroenterology video capsule endoscopy data set. Scientific Data **9**(1), 602 (2022)
11. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., et al.: Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021)
12. Le Floch, M., Wolf, F., McIntyre, L., Weinert, C., Palm, A., Volk, K., Herzog, P., Kirk, S.H., Steinhäuser, J.L., Stopp, C., et al.: Galar-a large multi-label video capsule endoscopy dataset. Scientific Data **12**(1), 828 (2025)
13. Werner, J., Gerum, C., Reiber, M., Nick, J., Bringmann, O.: Precise localization within the gi tract by combining classification of cnns and time-series analysis of hmms. In: International Workshop on Machine Learning in Medical Imaging. pp. 174–183. Springer (2023)
14. Werner, J., Bause, O., Oexle, J., Floch, M.L., Brinkmann, F., Hampe, J., Bringmann, O.: Seeing more with less: Video capsule endoscopy with multi-task learning. In: International Workshop on Applications of Medical AI. pp. 12–21. Springer (2025)
15. Boers, T.G., Fockens, K.N., van der Putten, J.A., Jaspers, T.J., Kusters, C.H., Jukema, J.B., Jong, M.R., Struyvenberg, M.R., de Groof, J., Bergman, J.J., et al.: Foundation models in gastrointestinal endoscopic ai: Impact of architecture, pretraining approach and data efficiency. Medical Image Analysis **98**, 103298 (2024)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Floch, M.L., Wolf, F., McIntyre, L., Herzog, P., Weinert, C., Palm, A., Volk, K., Kirk, S.H., Steinhäuser, J.L., Stopp, C., Geissler, M.E., Herzog, M., Sulk, S., Kather, J.N., Meining, A., Hann, A., Hampe, J., Herzog, N., Brinkmann, F.: Galar - a large multi-label video capsule endoscopy dataset. Figshare. https://doi.org/10.25452/figshare.plus.25304616.v2 (2 2025)
19. Charoen, A., Guo, A., Fangsaard, P., Taweechainaruemitr, S., Wiwatwattana, N., Charoenpong, T., Rich, H.: Rhode island gastroenterology video capsule endoscopy data set (Oct 2022). https://doi.org/10.6084/m9.figshare.c.6071216.v1
20. Jong, M.R., Boers, T.G., Fockens, K.N., Jukema, J.B., Kusters, C.H., Jaspers, T.J., van Heslinga, R.v.E., Slooter, F.C., Struyvenberg, M.R., Bisschops, R., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. Gastroenterology (2025)

21. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
22. Fein, M., Maroske, J., Fuchs, K.: Importance of duodenogastric reflux in gastro-oesophageal reflux disease. Journal of British Surgery **93**(12), 1475–1482 (2006)
23. Werner, J., Gerum, C., Nick, J., Le Floch, M., Brinkmann, F., Hampe, J., Bringmann, O.: Enhanced anomaly detection for capsule endoscopy using ensemble learning strategies. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1–7. IEEE (2025)
24. Melson, J., Trikudanathan, G., Dayyeh, B.K.A., Bhutani, M.S., Chandrasekhara, V., Jirapinyo, P., Krishnan, K., Kumta, N.A., Pannala, R., Parsi, M.A., et al.: Video capsule endoscopy. Gastrointestinal endoscopy **93**(4), 784–796 (2021)
25. Enns, R.A., Hookey, L., Armstrong, D., Bernstein, C.N., Heitman, S.J., Teshima, C., Leontiadis, G.I., Tse, F., Sadowski, D.: Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology **152**(3), 497–514 (2017)