Medico 2025: Visual Question Answering for Gastrointestinal Imaging

Sushant Gautam[1,3], Vajira Thambawita[1], Michael Riegler[2], Pål Halvorsen[1,3] and Steven Hicks[1,*]

[1]*SimulaMet - Simula Metropolitan Center for Digital Engineering, Oslo, Norway*
[2]*Simula Research Laboratory, Oslo, Norway*
[3]*OsloMet - Oslo Metropolitan University, Oslo, Norway*

Abstract

The Medico 2025 challenge addresses Visual Question Answering (VQA) for Gastrointestinal (GI) imaging, organized as part of the MediaEval tasks series. The challenge focuses on developing Explainable Artificial Intelligence (XAI) models that answer clinically relevant questions based on GI endoscopy images while providing interpretable justifications aligned with medical reasoning. It introduces two subtasks: (1) answering diverse types of visual questions using the Kvasir-VQA-x1 dataset, and (2) generating multimodal explanations to support clinical decision-making. The Kvasir-VQA-x1 dataset, created from 6,500 images and 159,549 complex question–answer (QA) pairs, serves as the benchmark for the challenge. By combining quantitative performance metrics and expert-reviewed explainability assessments, this task aims to advance trustworthy Artificial Intelligence (AI) in medical image analysis. Instructions, data access, and an updated guide for participation are available in the official competition repository: github.com/simula/MediaEval-Medico-2025

1. Introduction

Gastrointestinal (GI) diseases are among the most common and critical health concerns worldwide, with conditions like Colorectal Cancer (CRC) requiring early diagnosis and intervention [1, 2]. AI-driven decision support systems [3, 4] have shown potential in assisting clinicians with diagnosis, but a major challenge remains: explainability. While deep learning models can achieve high diagnostic accuracy, their "black-box" nature limits their adoption in clinical practice, where trust and interpretability are essential [5, 6]. After successfully organizing multiple Medico challenges at MediaEval in previous years, for the new edition[1] we propose the Medico 2025: *Visual Question Answering (with multimodal explanations) for Gastrointestinal Imaging*.

Medical VQA is a rapidly growing research area that combines computer vision and natural language processing to answer clinically relevant questions based on medical images [5]. However, existing VQA models often lack transparency, making it difficult for healthcare professionals to assess the reliability of AI-generated answers [5, 6]. To address this, the Medico 2025 challenge will focus on explainable VQA for GI imaging, encouraging participants to develop models that provide not only accurate answers but also clear justifications aligned with clinical reasoning.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

sushant@simula.no (S. Gautam); vajira@simula.no (V. Thambawita); michael@simula.no (M. Riegler); paalh@simula.no (P. Halvorsen); steven@simula.no (S. Hicks)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://multimediaeval.github.io/editions/2025/tasks/medico

The challenge provides a benchmark dataset of GI images, videos, and associated VQA annotations, enabling rigorous evaluation of AI models. By integrating multimodal data and explainability metrics, we aim to advance research in interpretable AI and increase the potential for clinical adoption.

We define two main subtasks for this year's challenge. Subtask 2 builds on Subtask 1, meaning Subtask 1 must be completed in order to participate in Subtask 2.

- **Subtask 1:** AI Performance on Medical Image Question Answering
  This subtask challenges participants to develop AI models that accurately interpret and respond to clinical questions based on GI images from the Kvasir-VQA-x1 dataset, which retains the original 6,500 images from Kvasir-VQA [7] but expands them to 159,549 QA pairs across multiple conditions and instruments. Questions fall into six categories: Yes/No, Single-Choice, Multiple-Choice, Color-Related, Location-Related, and Numerical Count, requiring models to process both visual and textual information.
  Performance will be assessed using Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1/2/L, and Metric for Evaluation of Translation with Explicit ORdering (METEOR).
- **Subtask 2:** Clinician-Oriented Multimodal Explanations in GI
  This subtask builds upon Subtask 1, requiring participants to justify their model's predictions using multiple complementary forms of reasoning. The goal is to generate rich, multimodal explanations that are transparent, understandable, and trustworthy to clinicians [8]. At a minimum, explanations must include a detailed textual narrative in clinical language that directly supports the predicted answer [9]. Participants are strongly encouraged to provide an accompanying visual explanation—such as a heatmap, segmentation mask, or bounding box—that clearly links to the textual reasoning and highlights the relevant finding [10, 11, 12]. Confidence scores, indicating the model's certainty, are optional but recommended.
  All outputs will be *human-evaluated* by domain experts and medical professionals, using predefined criteria for clarity, coherence between modalities, and medical relevance, to assess how well the outputs support clinical decision-making.

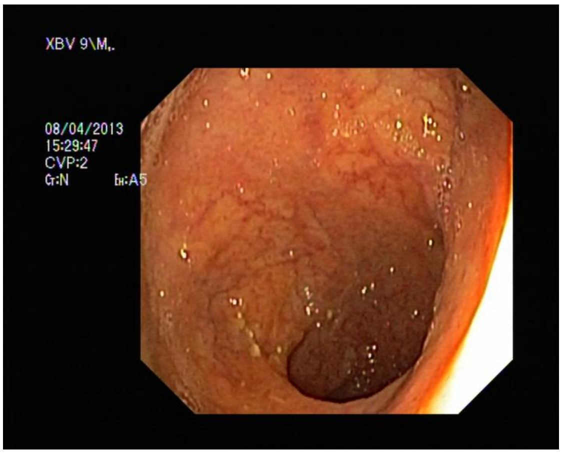
Medical AI systems must be both accurate and interpretable to be useful in clinical practice. While deep learning models have shown great potential in diagnosing GI conditions from medical images, their adoption remains limited due to a lack of transparency. Clinicians need to understand why an AI system makes a specific decision, especially when it comes to critical medical diagnoses. XAI methods aim to bridge this gap by providing justifications that align with clinical reasoning, improving trust, reliability, and ultimately patient outcomes.

This challenge builds on previous work in medical VQA, where AI models answer clinically relevant questions based on GI images. However, traditional VQA models often provide answers without explanations, making it difficult for medical professionals to assess their validity. By incorporating explainability into the task, we encourage the development of models that not only provide accurate responses but also offer meaningful insights into their decision-making process. This will help ensure that AI systems can be safely integrated into clinical workflows, assisting rather than replacing human expertise.

2. Data

The Medico 2025 challenge builds on the **Kvasir-VQA-x1** dataset [13], a substantial extension of the original Kvasir-VQA [7]. It comprises **6,500** GI endoscopic images from HyperKvasir [14] and Kvasir-Instrument [15], paired with **159,549** QA pairs stratified by reasoning complexity.

Table 1
An example image with one representative question–answer pair from each complexity level from the Kvasir-VQA-x1 dataset. Each image in the dataset may have multiple QA pairs at every level.



Complexity	Question	Answer	Question Class
1	Which anatomical landmark is visible in the image?	No identifiable anatomical landmark present	landmark_location
2	What procedure is depicted in the image and what colors are associated with the abnormality?	Evidence of colonoscopy findings with pink and red mucosal lesions	procedure_type, abnormality_color
3	Are there any anatomical landmarks visible, what type of polyps are present, and what colors are the observed abnormalities?	No anatomical landmarks identified, no polyps observed, and multiple abnormalities with pink and red coloration.	landmark_presence, polyp_type, abnormality_color

2.1. Dataset Composition

Each image is paired with multiple QA entries generated by merging one to three atomic QA pairs using the Qwen3-30B-A3B model [16]. The resulting natural-language questions are fluently phrased, and each entry is annotated with a `complexity` score (ranging from 1 to 3) and a `question_class` label specifying its clinical category. Examples of these classes include `polyp_type`, `instrument_presence`, and `finding_count`.

2.2. Question Complexity and Clinical Categorization

The dataset supports stratified evaluation by QA complexity:

- **Level 1**: Questions derived from a single atomic QA (approximately 34.4%).
- **Level 2**: Reasoning over two merged atomic QA (32.8%).
- **Level 3**: Synthesis across three atomic QA (32.8%).

Each QA pair is assigned one or more `question_class` labels to support fine-grained analysis across clinical categories such as pathology, anatomical localization, procedural context, and visual findings.

Public Availability and Format

Kvasir-VQA-x1 is hosted at https://huggingface.co/datasets/SimulaMet/Kvasir-VQA-x1. The dataset includes `img_id` (same as in Kvasir-VQA [7]), `complexity`, `question`, `answer`, `original` (atomic QA components), and `question_class`. It is split into training and testing subsets for reproducible experimentation. While only the original images are released, we encourage applying weak augmentations (e.g., rotation, color jitter, crop) when fine-tuning models with the dataset.

3. Evaluation

The Medico 2025 challenge evaluates both the **accuracy** and **clinical interpretability** of medical VQA models, emphasizing not only correct answers but also their **relevance and explanatory quality** in the context of GI diagnostics.

3.1. Subtask 1: GI Question Answering

This subtask evaluates how effectively models answer clinically relevant GI questions from medical images, emphasizing both **predictive accuracy** and **reasoning depth**. Performance is measured using language quality metrics—BLEU, ROUGE-1/2/L, and METEOR—to assess alignment with reference responses. Evaluation is conducted in two settings: an *original* setting with clean images and a *transformed* setting that applies augmentations for robustness testing. Criteria include accuracy, relevance, and medical correctness.

Evaluation is stratified across three levels: (1) *overall performance*, aggregating scores across all categories and complexities; (2) *category-level analysis* over 18 question types (e.g., polyp type, instrument presence), with visualizations such as radar plots and rank-normalized heatmaps; and (3) *complexity-level evaluation*, distinguishing between factual (Level 1), moderately inferential (Level 2), and higher-order reasoning prompts (Level 3).

This structured, multi-dimensional evaluation framework provides a comprehensive assessment of both correctness and clinical reasoning, which is critical for robust deployment in medical settings.

3.2. Subtask 2: Clinician-Oriented Explanation Quality Assessment

This subtask extends Subtask 1 by requiring participants to justify their predictions through detailed, multimodal explanations. The objective is to move beyond providing an answer and produce outputs that are transparent, clinically relevant, and aligned with the model's own reasoning.

Evaluation in Subtask 2 will be conducted entirely by human experts and medical professionals. Automated metrics from Subtask 1 (e.g., BLEU, ROUGE-1/2/L, METEOR) are used only to assess answer correctness; explanation quality will be judged through expert review according to the criteria below.

Each explanation must combine:

- **Textual Explanation (Mandatory):** A detailed, clinician-oriented narrative that justifies the predicted answer using multiple aspects.
- **Visual Explanation (Optional but Highly Encouraged):** A supporting visual modality—such as a Grad-CAM heatmap, segmentation mask, or bounding box—that highlights the region(s) referenced in the textual explanation. Visuals must clearly link to and reinforce the textual reasoning.

- **Confidence Score (Optional):** A scalar in [0, 1] indicating the model's certainty in its prediction, derived from softmax probabilities, calibrated uncertainty, or Bayesian methods.

Submissions must follow a structured JSON and expert reviewers will rate submissions based on (but not limited to) the following criteria:

- **Clarity:** Ease of understanding for a clinician.
- **Coherence:** Logical consistency between visual and textual components.
- **Medical Relevance:** Consistency with established clinical knowledge.
- **Visual Alignment:** Whether visual elements accurately highlight the relevant findings.

By combining accurate predictions with interpretable, clinically grounded justifications, this subtask aims to promote AI systems that can be meaningfully integrated into real-world diagnostic workflows.

4. Discussion and Outlook

The Medico 2025 challenge marks an important step toward bridging the gap between powerful deep learning models and their practical adoption in clinical settings. By focusing on explainable VQA for GI imaging, this task promotes the development of interpretable AI models that not only generate accurate responses but also provide transparent justifications aligned with medical reasoning.

Participants are encouraged to innovate beyond traditional accuracy metrics and embrace multimodal explainability as a core component of their solutions. The availability of the Kvasir-VQA-x1 dataset, tailored for this task, will support reproducible research and enable robust benchmarking.

Looking ahead, we anticipate that methods developed for Medico 2025 will inspire broader applications of explainable AI in other medical domains. By fostering interdisciplinary collaboration between the AI and medical communities, this challenge aims to pave the way for clinically viable AI tools that are both trusted and actionable in real-world healthcare scenarios.

References

[1] A. Singh, Global burden of five major types of gastrointestinal cancer, Gastroenterology Review/Przegląd Gastroenterologiczny 19 (2024).

[2] R. Wang, Z. Li, S. Liu, D. Zhang, Global, regional, and national burden of 10 digestive diseases in 204 countries and territories from 1990 to 2019, Frontiers in public health 11 (2023) 1061453.

[3] H. Ali, M. A. Muzammil, D. S. Dahiya, F. Ali, S. Yasin, W. Hanif, M. K. Gangwani, M. Aziz, M. Khalaf, D. Basuli, et al., Artificial intelligence in gastrointestinal endoscopy: a comprehensive review, Annals of gastroenterology 37 (2024) 133.

[4] M. A. Berbís, J. Aneiros-Fernández, F. J. M. Olivares, E. Nava, A. Luna, Role of artificial intelligence in multidisciplinary imaging diagnosis of gastrointestinal diseases, World journal of gastroenterology 27 (2021) 4395.

[5] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches, Eur. J. Radiol. 162 (2023). doi:10.1016/j.ejrad.2023.110786.

[6] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111. doi:10.1016/j.compbiomed.2021.105111.

[7] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications, 2024, pp. 3–12.

[8] D. Muhammad, M. Bendechache, Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis, Comput. Struct. Biotechnol. J. 24 (2024) 542–560. doi:10.1016/j.csbj.2024.08.005.

[9] X. Gai, C. Zhou, J. Liu, Y. F. (xn-27q. xn-6xw.), J. Wu, Z. Liu, MedThink: A Rationale-Guided Framework for Explaining Medical Visual Question Answering, ACL Anthology (2025) 7438–7450. doi:10.18653/v1/2025.findings-naacl.415.

[10] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, IEEE Computer Society, 2018. doi:10.1109/CVPR.2018.00915.

[11] A. M. Storås, M. Dreyer, F. Pahde, S. Lapuschkin, W. Samek, P. Halvorsen, T. de Lange, Y. Mori, A. Hann, T. M. Berzin, S. Parasa, M. A. Riegler, Exploring the clinical value of concept-based AI explanations in gastrointestinal disease detection, Sci. Rep. 15 (2025) 1–11. doi:10.1038/s41598-025-14408-y.

[12] F. Dahan, J. H. Shah, R. Saleem, M. Hasnain, M. Afzal, T. M. Alfakih, A hybrid XAI-driven deep learning framework for robust GI tract disease diagnosis, Sci. Rep. 15 (2025) 1–18. doi:10.1038/s41598-025-07690-3.

[13] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.

[14] 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, et al., Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific data 7 (2020) 283.

[15] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. De Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27, Springer, 2021, pp. 218–229.

[16] A. Yang, A. Li, B. Yang, et al., Qwen3 Technical Report, arXiv (2025). doi:10.48550/arXiv.2505.09388. arXiv:2505.09388.