Overview of ImageCLEFmedical 2026 – Visual Question Answering with Explainable and Safety-Aware Reasoning for the Gastrointestinal Tract

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

This paper provides an overview of the fourth edition of the Medical Visual Question Answering for the Gastrointestinal Tract (MedVQA-GI) challenge, hosted at ImageCLEF 2026. Building on insights gained from the previous three editions, this year's challenge shifts focus from data and image generation toward trustworthy multimodal reasoning in realistic clinical settings. The challenge comprises two tasks: (1) clinically relevant Visual Question Answering (VQA) over gastrointestinal (GI) endoscopy images, and (2) multimodal explainability and safety-aware reasoning, where models must justify their answers using textual explanations and visual evidence while adhering to clinical safety principles. A dedicated behavioral safety evaluation examines undesirable model behaviors such as overconfidence, hallucinations, and misleading explanations. The challenge builds on Kvasir-VQA-x1, an expanded version of the Kvasir-VQA dataset comprising more than 150,000 question-answer pairs derived from GI endoscopy images. Submissions are evaluated using a combination of automatic text-generation metrics, expert-based assessment, and behavioral safety evaluation. This paper details the tasks, data, and evaluation methodology, summarizes the participating systems, and reports the results of Task 1. The competition repository is at: github.com/simula/ImageCLEFmed-MEDVQA-GI-2026.

Keywords

Visual question answering, Explainability, Trustworthy and safe AI, Endoscopy, Machine learning

1. Introduction

The fourth edition of the Medical Visual Question Answering for the Gastrointestinal Tract (MedVQA-GI) challenge at ImageCLEF continues our focus on advanced image-based machine learning for gastrointestinal (GI) diagnostics. Building on insights and outcomes from the previous three editions, this year we shift focus from data and image generation toward trustworthy multimodal reasoning in realistic clinical settings. Models must not only answer clinically relevant questions from GI endoscopy images, but also justify their answers in a manner aligned with medical reasoning while adhering to clinical safety principles. The challenge explicitly addresses emerging concerns related to overconfidence, hallucinations, and misleading explanations in medical AI systems.

Machine learning has long been applied to support lesion detection in gastrointestinal (GI) images [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Historically, most efforts have centered on identifying abnormalities such as polyps in image or video data [12, 13, 14, 15, 16, 17, 18], and multiple shared tasks have advanced the field through organized benchmarking [19, 20, 21, 22, 23]. As such systems move closer to clinical deployment, attention has shifted from raw predictive accuracy toward trustworthiness: models should be interpretable, well-calibrated, and safe, avoiding overconfident or misleading outputs that could distort clinical decision-making. To reflect these concerns, and by integrating behavioral safety evaluation and retrieval-augmented reasoning, this year's MedVQA-GI challenge centers on

CLEF 2026 Working Notes, September 2026, Jena, Germany
*Corresponding author.
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)

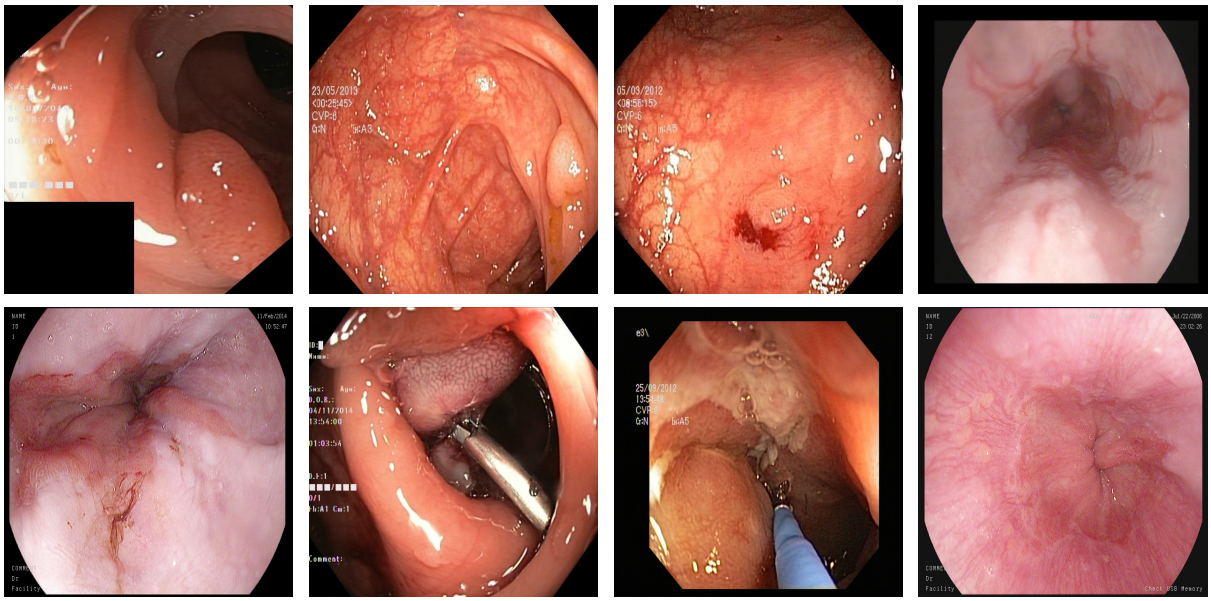

Figure 1: Examples from the development dataset provided by the challenge organizers. The images represent various types of findings and contexts.

clinically relevant VQA paired with explainable, safety-aware reasoning. All data and supporting code are available in our public repository[1].

The remainder of this paper is organized as follows. First, we describe the dataset and its structure. Then, we present the two challenge tasks along with the evaluation methodology, followed by a description of the submission system. Finally, we conclude with the goals and expected outcomes of this year's challenge.

2. Dataset

The data used in this year's challenge is based on Kvasir-VQA-x1 [24, 25], an expanded version of the Kvasir-VQA [26] dataset, which itself builds on the publicly available HyperKvasir [27] dataset. Kvasir-VQA-x1 comprises more than 150,000 question-answer pairs derived from gastrointestinal (GI) endoscopy images covering a wide range of anatomical sites and pathological findings. Crucially, the dataset supports both answer prediction and explainability evaluation, making it suitable for clinically relevant VQA as well as the generation and assessment of multimodal justifications. Examples from the dataset can be seen in Figure 1. The development dataset is publicly available[2].

The questions span multiple types and reflect clinically relevant scenarios, ranging from diagnostic and descriptive queries to questions requiring classification, reasoning, spatial localization, and attribute recognition. This diversity helps models generalize to real-world diagnostic tasks. As in previous editions, the test set is drawn from a separate, unreleased collection of GI images sampled from sources not included in the development set. This ensures the test data is distributionally distinct and unseen, allowing us to better evaluate the generalization performance of participating systems under realistic conditions.

3. Tasks and Evaluation

This year, MedVQA-GI is made up of two tasks: answering clinically relevant questions from GI endoscopy images, and producing explainable, safety-aware justifications for those answers. The

[1]https://github.com/simula/ImageCLEFmed-MEDVQA-GI-2026
[2]https://datasets.simula.no/kvasir-vqa-x1

evaluation differs for each task and combines objective metrics with expert-based assessment.

3.1. Task 1: Clinically Relevant Visual Question Answering

This task asks participants to develop algorithms that accurately answer clinically relevant questions based on gastrointestinal (GI) endoscopy images from the Kvasir-VQA-x1 dataset [24, 25]. Models must correctly interpret both the visual findings and the intent of the question to produce concise and medically sound answers, with a focus on the diagnostic and descriptive questions commonly encountered in clinical practice. Performance is assessed using automatic metrics that capture both answer correctness and language quality:

METEOR [28] Evaluates text generation by aligning predicted and reference outputs based on exact, stem, synonym, and paraphrase matches.

ROUGE (1/2/L) [29] A set of metrics for comparing overlapping n-grams between generated and reference texts. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L captures the longest common subsequence.

BLEU [30] Measures n-gram precision between generated and reference texts, with a brevity penalty to penalize overly short outputs.

In addition to these lexical metrics, we complement the evaluation with an LLM-as-a-judge protocol, in which a large language model (Qwen3-27B) rates each generated answer against the reference along five clinically motivated dimensions (correctness, faithfulness, clinical relevance, clarity, and completeness) on a 0–10 scale. This provides a semantic, format-tolerant view of answer quality that is less sensitive to surface wording than the lexical metrics.

3.2. Task 2: Multimodal Explainability and Safety-Aware Reasoning

This task extends standard VQA by requiring participants to generate coherent multimodal justifications that combine textual explanations with visual evidence drawn from the image. Explanations should reflect established medical reasoning and remain factual, consistent, and clinically safe. In addition to producing an answer, models are asked to provide a clinician-oriented textual explanation referencing relevant visual cues (location, morphology, color, size, vascular pattern, and similar), and, where possible, visual grounding artifacts such as heatmaps, segmentation masks, or bounding boxes that localize the evidence supporting the answer.

To assess model performance, we combine automatic measures with expert-based assessment of explanation quality, interpretability, clinical plausibility, and safety. A dedicated behavioral safety evaluation examines whether generated answers and explanations exhibit problematic behaviors such as hallucination, overconfidence, or unsafe clinical recommendations, optionally drawing on model-reported confidence or uncertainty estimates for calibration analysis. An online leaderboard provides rapid evaluation feedback to nurture a competitive environment.

4. Submission System

Submissions follow a different workflow for each task. For Task 1, models are submitted and validated through the `medvqa` Python package[3] against the official system, ensuring fair comparison and reproducibility. Participants place their trained model in a gated Hugging Face repository together with a standalone `submission_task1.py` script, validate it locally with `medvqa validate -competition=gi-2026 -task=1`, and then register an official submission with `medvqa validate_and_submit -competition=gi-2026 -task=1`. Once submitted, the participant's username, task, and submission time appear on the public portal[4], which also serves as the registered leaderboard.

[3] https://pypi.org/project/medvqa/

[4] https://simulamet-medvqa-gi-2026.hf.space

Table 1
An overview of the submissions to each task at MedVQA-GI.

	MedVQA 2023	MedVQA 2024	MedVQA 2025	MedVQA 2026
# Registrations	26	22	31	26
# Task Participation	8	2	5	8
# Paper Submissions	6	2	5	5

For Task 2, participants submit their results directly to the organizers by email as a ZIP archive. The archive contains a `submission_task2.jsonl` file with one JSON object per validation item (including the answer, a clinician-oriented textual explanation, an optional list of visual-explanation artifacts, and an optional confidence score in $[0, 1]$), together with a `submission_task2.py` file holding team metadata. Submissions are evaluated by the organizers and final results are presented at the workshop.

5. Participation

In total, 26 teams registered for the challenge. Eight teams submitted runs to Task 1, and five of these also submitted working notes papers [31, 32, 33, 34, 35]; the remaining three participated on the leaderboard only. Table 1 shows an overview of the participation this year alongside the numbers from previous editions. Three of the five (IReL, CS_Morgan Lab, and CARES) additionally developed explainability and safety-aware systems for Task 2. As in previous years, many who registered did not go on to submit, which remains a common pattern. Consistent with the more demanding, trustworthiness-oriented focus of this edition, the teams that did submit engaged deeply with the clinically grounded and safety-aware aspects of the challenge.

6. Results

Eight teams submitted runs to Task 1. Table 2 reports the lexical metrics and Table 3 the LLM-as-a-judge scores, in both cases on the public (Kvasir-VQA-test) and private test sets. Five teams accompanied their submissions with working notes, and we summarize their approaches below, ordered by public ROUGE-1. The remaining three participated on the leaderboard only: OmniMed (a masked-diffusion Qwen2.5-VL variant), Something Creative (a BM25 and ConceptCLIP retrieval pipeline with reciprocal-rank fusion), and SR0112 (a BiomedCLIP retrieval baseline). As noted above, the Task 2 explainability and safety evaluation is conducted separately, with results presented at the workshop.

The most striking pattern is a sharp reversal between the two test sets. Systems that topped the public leaderboard by fine-tuning or post-processing toward the compact, label-style public answers (for example, CARES and the RAG systems, with public ROUGE-1 above 0.80) collapse on the private set, whose references are written as full natural-language clinical sentences. Conversely, a retrieval-only system that returns training answers verbatim (Something Creative) achieves the best private scores, and the lightweight IReL model, which was not tuned to the public label format, generalizes best among the other teams on the private set. The LLM-as-a-judge scores are more robust to this format mismatch than the lexical metrics (faithfulness in particular remains high across both sets), but they too decline on the private set as question complexity increases. We return to these observations in the discussion.

6.1. Team CARES

Team CARES [34] achieved the top public-leaderboard scores (public ROUGE-1 of 0.89) using a training-free, inference-time reliability layer built on a frozen Florence-2-base model with a public LoRA adapter, without fine-tuning the backbone. For Task 1, a deterministic concise-normaliser maps the model's verbose outputs into canonical answer forms (procedure type, abnormality class, colour codes, and similar). For Task 2, they add a calibrated per-(aspect, answer) reliability table fit by empirical-Bayes

Table 2
Task 1 lexical metrics on the public (Kvasir-VQA-test) and private test sets. B-1/B-2: BLEU-1/BLEU-2; R-1/R-2/R-L: ROUGE-1/ROUGE-2/ROUGE-L; MET: METEOR. For all metrics, higher is better (↑); the best value in each column on the private test set is shown in bold. Teams are ordered by public ROUGE-1; † marks the three teams that submitted runs without accompanying working notes.

Team	Set	B-1 ↑	B-2 ↑	R-1 ↑	R-2 ↑	R-L ↑	MET ↑
CARES	Public	0.7480	0.6350	0.8932	0.1325	0.8882	0.4979
	Private	0.0005	0.0002	0.1069	0.0125	0.1054	0.0382
TUW_AIR22	Public	0.7300	0.5600	0.8400	0.0900	0.8200	0.4600
	Private	0.0000	0.0000	0.1300	0.0100	0.1300	0.0400
UIT_NEWRON	Public	0.6397	0.5468	0.8016	0.0914	0.8003	0.4354
	Private	0.0005	0.0002	0.1369	0.0067	0.1354	0.0465
OmniMed†	Public	0.2549	0.1910	0.6641	0.0287	0.6639	0.3272
	Private	0.0001	0.0000	0.1382	0.0012	0.1379	0.0435
IReL, IIT (BHU)	Public	0.1748	0.0996	0.6011	0.0581	0.5987	0.3744
	Private	0.2500	0.2000	0.3900	0.2500	0.3500	0.4800
CS_Morgan Lab	Public	0.1918	0.1185	0.5160	0.0095	0.5052	0.2594
	Private	0.0035	0.0022	0.1987	0.0232	0.1916	0.0864
Something Creative†	Public	0.0300	0.0000	0.0500	0.0000	0.0500	0.0500
	Private	**0.5200**	**0.4000**	**0.5700**	**0.3600**	**0.5300**	**0.5200**
SR0112†	Public	0.0032	0.0000	0.0178	0.0000	0.0180	0.0267
	Private	0.1537	0.0721	0.1476	0.0322	0.1243	0.1211

Table 3
Task 1 LLM-as-a-judge scores (Qwen3-27B, five dimensions on a 0–10 scale) on the public and private test sets. Corr.: correctness; Faith.: faithfulness; Clin.: clinical relevance; Clar.: clarity; Compl.: completeness. For all dimensions, higher is better (↑); the best value in each column on the private test set is shown in bold. Teams are ordered as in Table 2; † marks teams without accompanying working notes.

Team	Set	Corr. ↑	Faith. ↑	Clin. ↑	Clar. ↑	Compl. ↑
CARES	Public	8.82	9.61	9.50	9.83	9.68
	Private	5.50	9.05	5.12	7.46	5.21
TUW_AIR22	Public	8.35	9.59	9.43	9.85	9.33
	Private	5.06	8.71	5.14	7.43	5.01
UIT_NEWRON	Public	7.92	9.64	8.92	9.94	8.87
	Private	5.11	8.68	5.00	7.26	5.14
OmniMed†	Public	7.16	8.43	7.92	9.28	7.76
	Private	4.97	8.35	4.47	6.96	4.49
IReL, IIT (BHU)	Public	7.49	8.25	7.81	7.48	8.40
	Private	**6.49**	4.95	6.12	6.08	8.22
CS_Morgan Lab	Public	7.19	8.88	8.16	9.77	8.11
	Private	5.35	**9.15**	5.03	7.59	5.06
Something Creative†	Public	3.34	5.23	5.01	8.74	4.52
	Private	6.12	8.07	**7.71**	**9.52**	**8.26**
SR0112†	Public	2.58	4.27	2.90	6.01	3.54
	Private	0.89	2.59	1.41	6.79	1.19

shrinkage, a self-probing explanation template gated to avoid attributing findings to negative cases, referring-expression segmentation for visual grounding, and an assert/hedge/abstain safety policy with a Clopper–Pearson selective-risk guarantee. Their heavy reliance on public-format normalisation, however, caused a large drop on the private test.

6.2. Team TUW_AIR22

Team TUW_AIR22 [35] participated in Task 1 with a retrieval-augmented generation (RAG) approach, submitting several configurations. Their strongest run couples a Qwen2.5-VL-7B model using the SimulaMet fine-tuned LoRA adapter with selective BM25 retrieval that supplies structured clinical

evidence for anatomy, abnormality, and procedure questions while disabling retrieval for purely visual questions such as colour and location; a hybrid variant additionally retrieves visually similar images with ColQwen2 embeddings. Extensive post-processing normalises outputs to canonical forms, which yielded strong public performance (ROUGE-1 of 0.84) but limited generalisation to the private set's natural-language references.

6.3. Team UIT_NEWRON

Team UIT_NEWRON [33] fine-tuned Qwen2.5-VL-7B-Instruct via 4-bit QLoRA in an NF4 configuration and paired it with an extensive rule-based post-processor. Their pipeline groups all questions for the same image and issues targeted context queries (polyp presence, finding type, procedure) to build a per-image context cache before applying deterministic rules, enforcing consistency across related answers. A post-competition failure analysis identified branch-ordering and context-gating fragilities under distribution shift, which they subsequently diagnosed and corrected. The system reached a public ROUGE-1 of 0.80 but, like other post-processing-heavy systems, degraded sharply on the private test.

6.4. Team IReL, IIT (BHU)

Team IReL, IIT (BHU) [31] took a deliberately lightweight route, combining a frozen InceptionV3 image encoder, a Bio_ClinicalBERT question encoder, and a BioGPT decoder joined by a learned prefix-fusion layer, with beam search under no-repeat n-gram constraints. For Task 2, they extend the framework with Grad-CAM-based visual grounding, token-level confidence estimation, Monte Carlo dropout for uncertainty quantification, and semantic alignment analysis. Because the model was not tuned to the compact public answer format, it generalized notably better than the fine-tuned VLMs on the private set, where it recorded the second-highest private METEOR (0.48) of all teams, behind only the Something Creative retrieval baseline.

6.5. Team CS_Morgan Lab

Team CS_Morgan Lab [32] fine-tuned Qwen2.5-VL-7B-Instruct with 4-bit QLoRA, continuing from the SimulaMet Kvasir-VQA-x1 adapter, and combined it with strict label-only, question-type-aware prompting and a canonical-answer normaliser. This design strongly suppressed hallucination: its LLM-judged faithfulness stayed high on both sets and was the highest of all teams on the private set (9.15, versus 8.88 on the public set), though the label-only constraint limited lexical overlap with the private set's descriptive references. For Task 2, they fine-tuned MedGemma-4B-IT with QLoRA to produce a structured four-field response (answer, justification, confidence, and safety note) conditioned on the Task 1 prediction, and integrated a SAM segmenter fine-tuned on Kvasir-SEG to supply polyp masks and bounding boxes, gated to suppress localization for negative findings.

7. Discussion

Across Task 1, most teams built on large vision-language models adapted with parameter-efficient QLoRA fine-tuning, with Qwen2.5-VL-7B the most common backbone and Florence-2 also represented, and several teams paired the model with deterministic rule-based post-processing that maps free-form generations into the concise clinical answer format. The range of approaches was unusually diverse this year, also including retrieval-augmented generation (TUW_AIR22), pure retrieval baselines over biomedical embeddings (SR0112 and Something Creative), and a masked-diffusion language model with bidirectional attention (OmniMed).

The dominant finding is a large and systematic gap between the public and private test sets. On the public split, the leading systems exceeded a ROUGE-1 of 0.80 (up to 0.89), yet the same systems fell below 0.15 on the private split. The teams' own analyses attribute this to a difference in answer annotation style rather than model quality: the public references are compact labels (for example,

"none" or "colonoscopy"), whereas the private references are full natural-language clinical sentences. Systems that fine-tuned or post-processed toward the compact public format could not reproduce the descriptive private answers, while a retrieval baseline that returns training answers verbatim rose from near the bottom of the public leaderboard to the top of the private one. This underscores a limitation of surface-level lexical metrics for this task, as they conflate answer correctness with formatting conventions. The LLM-as-a-judge scores were more robust: faithfulness remained high (often above 8 out of 10) even where lexical scores collapsed, indicating that models were generally not hallucinating but rather mismatching the expected answer form. Even so, judge scores on all systems still declined on multi-aspect (dual- and triple-aspect) questions, which remain the hardest setting. A clear lesson for future editions is to align the annotation style of the public and private references, or to report complementary metrics that are robust to formatting.

In response to this year's emphasis on trustworthy reasoning, several teams introduced techniques that were largely absent from previous editions, including confidence calibration, structured and self-probing explanation templates, uncertainty quantification, visual grounding via Grad-CAM or segmentation, and explicit assert/hedge/abstain safety policies; the corresponding Task 2 explainability and safety results are presented at the workshop. Although the challenge attracted a healthy number of registrations, the number of completed submissions remained modest, consistent with previous editions and reflecting the technical and resource demands of the tasks. The trustworthiness focus of Task 2, in particular, asks participants to move beyond raw predictive accuracy, and the submissions this year suggest that the community is beginning to engage seriously with explainability and safety in medical VQA.

8. Conclusion

This paper presented the 2026 edition of the MedVQA-GI challenge, held as part of ImageCLEF. The fourth edition shifts focus from data and image generation toward trustworthy multimodal reasoning, comprising two tasks: clinically relevant VQA and multimodal explainability with safety-aware reasoning. By combining automatic metrics, expert assessment, and a dedicated behavioral safety evaluation over the large-scale Kvasir-VQA-x1 dataset, the challenge aims to promote the development of reliable, interpretable, and clinically robust VQA systems for endoscopy. Five teams submitted working notes to this edition; we summarize their approaches and report the Task 1 results, while the Task 2 explainability and safety evaluation is finalized separately and presented at the workshop.

Declaration on Generative AI

During the preparation of this work, the authors used Claude for grammar and spelling checks, paraphrasing and rewording, and improving the writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Hassan, M. Spadaccini, A. Iannone, R. Maselli, M. Jovani, V. T. Chandrasekar, G. Antonelli, H. Yu, M. Areia, M. Dinis-Ribeiro, et al., Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis, Gastrointestinal endoscopy 93 (2021) 77–85.

[2] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, J. Wong, P. C. De Groen, Classification of ulcerative colitis severity in colonoscopy videos using cnn, in: Proceedings of the ACM International Conference on Information Management and Engineering (ACM ICIME), 2017, pp. 139–144. doi:https://doi.org/10.1145/3149572.3149613.

[3] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, J. Lundin, Deep learning based tissue analysis predicts outcome in colorectal cancer,

Scientific Reports 8 (2018) 3395. URL: http://dx.doi.org/10.1038/s41598-018-21758-3. doi:https://doi.org/10.1038/s41598-018-21758-3.

[4] Y. Mori, S.-e. Kudo, M. Misawa, Y. Saito, H. Ikematsu, K. Hotta, K. Ohtsuka, F. Urushibara, S. Kataoka, Y. Ogawa, Y. Maeda, K. Takeda, H. Nakamura, K. Ichimasa, T. Kudo, T. Hayashi, K. Wakamura, F. Ishida, H. Inoue, H. Itoh, M. Oda, K. Mori, Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study, Annals of Internal Medicine 169 (2018) 357–366. doi:https://doi.org/10.7326/M18-0249.

[5] K. Pogorelov, S. L. Eskeland, T. de Lange, C. Griwodz, K. R. Randel, H. K. Stensland, D.-T. Dang-Nguyen, C. Spampinato, D. Johansen, M. Riegler, P. Halvorsen, A holistic multimedia system for gastrointestinal tract disease detection, in: Proceedings of the ACM on Multimedia Systems Conference (MMSYS), 2017, pp. 112–123. doi:https://doi.org/10.1145/3193740.

[6] J. Silva, A. Histace, O. Romain, X. Dray, 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 (2014) 283–293. doi:https://doi.org/10.1007/s11548-013-0926-3.

[7] V. L. Thambawita, D. Jha, H. L. Hammer, H. D. Johansen, D. Johansen, P. Halvorsen, M. Riegler, An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification, ACM Transactions on Computing for Healthcare (2020).

[8] D. Jha, M. Riegler, D. Johansen, P. Halvorsen, H. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: Proceeding of the International Symposium on Computer Based Medical Systems (CBMS), 2020.

[9] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, A. Histace, Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis, in: Proceedings of Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE CLIP), volume 10550, Springer, 2017, pp. 29–41.

[10] K. Pogorelov, M. Riegler, P. Halvorsen, P. T. Schmidt, C. Griwodz, D. Johansen, S. L. Eskeland, T. de Lange, Gpu-accelerated real-time gastrointestinal diseases detection, in: Proceedings of the International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2016, pp. 185–190. doi:https://doi.org/10.1109/CBMS.2016.63.

[11] M. Riegler, K. Pogorelov, P. Halvorsen, T. de Lange, C. Griwodz, P. T. Schmidt, S. L. Eskeland, D. Johansen, EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies, in: Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), 2016, pp. 1–6. doi:https://doi.org/10.1109/CBMI.2016.7500257.

[12] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, P. C. De Groen, Polyp-alert: Near real-time feedback during colonoscopy, Computer Methods and Programs in Biomedicine 120 (2015) 164–179. doi:https://doi.org/10.1016/j.cmpb.2015.04.002.

[13] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: Proceedings of the International Symposium on Multimedia (ISM), 2019, pp. 225–230. doi:https://doi.org/10.1109/ISM46123.2019.00049.

[14] J. Bernal, A. Histace, M. Masana, Q. Angermann, C. Sánchez-Montes, C. Rodriguez, M. Hammami, A. Garcia-Rodriguez, H. Córdova, O. Romain, G. Fernández-Esparrach, X. Dray, J. Sanchez, Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases, in: Proceedings of Computer Assisted Radiology and Surgery (CARS), 2018. doi:https://hal.archives-ouvertes.fr/hal-01846141.

[15] Y. Guo, J. Bernal, B. J Matuszewski, Polyp segmentation with fully convolutional deep neural networks—extended evaluation study, Journal of Imaging 6 (2020) 69.

[16] M. Min, S. Su, W. He, Y. Bi, Z. Ma, Y. Liu, Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology, Scientific reports 9 (2019) 2881. doi:https://doi.org/10.1038/s41598-019-39416-7.

[17] N. M. Ghatwary, X. Ye, M. Zolgharni, Esophageal abnormality detection using densenet based

faster r-cnn with gabor features, IEEE Access 7 (2019) 84374–84385. doi:https://doi.org/10.1109/ACCESS.2019.2925585.

[18] S. Shah, N. Park, N. E. H. Chehade, A. Chahine, M. Monachese, A. Tiritilli, Z. Moosvi, R. Ortizo, J. Samarasena, Effect of computer-aided colonoscopy on adenoma miss rates and polyp detection: a systematic review and meta-analysis, Journal of Gastroenterology and Hepatology 38 (2023) 162–176.

[19] S. Hicks, M. Riegler, P. Smedsrud, T. B. Haugen, K. R. Randel, K. Pogorelov, H. K. Stensland, D.-T. Dang-Nguyen, M. Lux, A. Petlund, T. de Lange, P. T. Schmidt, P. Halvorsen, Acm multimedia biomedia 2019 grand challenge overview, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2019, pp. 2563–2567. doi:https://doi.org/10.1145/3343031.3356058.

[20] K. Pogorelov, M. Riegler, P. Halvorsen, S. A. Hicks, K. R. Randel, D.-T. Dang-Nguyen, M. Lux, O. Ostroukhova, T. De Lange, Medico multimedia task at mediaeval 2018, in: Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval), 2018.

[21] M. Riegler, K. Pogorelov, P. Halvorsen, K. Randel, S. Eskeland, D.-T. Dang-Nguyen, M. Lux, C. Griwodz, C. Spampinato, T. de Lange, Multimedia for medicine: the medico task at mediaeval 2017, in: Proceeding of the MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop (MediaEval), 2017.

[22] J. Bernal, H. Aymeric, Miccai endoscopic vision challenge polyp detection and segmentation, https://endovissub2017-giana.grand-challenge.org/home/, 2017. Accessed: 2017-12-11.

[23] S. Hicks, M. Riegler, P. Smedsrud, T. B. Haugen, K. R. Randel, K. Pogorelov, H. K. Stensland, D.-T. Dang-Nguyen, M. Lux, A. Petlund, T. de Lange, P. T. Schmidt, P. Halvorsen, Acm multimedia biomedia 2019 grand challenge overview, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2563–2567. URL: https://doi.org/10.1145/3343031.3356058. doi:10.1145/3343031.3356058.

[24] S. Gautam, M. Riegler, P. Halvorsen, Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy, in: Data Engineering in Medical Imaging, Springer, Cham, 2025. doi:10.1007/978-3-032-08009-7_6.

[25] 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.

[26] S. Gautam, A. 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 (VLM4Bio), ACM, 2024, p. 10 pages. doi:10.1145/3689096.3689458.

[27] 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). doi:10.1038/s41597-020-00622-y.

[28] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909.

[29] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[30] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[31] K. Tewari, S. Pal, From perception to explanation: A multimodal framework for gi visual question answering, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[32] R. N. Chowdhury, O. U. Islam, E. P. O. Oluwafemi, M. Hoque, E. F. Akor, M. M. Rahman, Parameter-efficient vision–language fine-tuning for gastrointestinal endoscopy vqa with structured explainability and safety-oriented reasoning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[33] T. N. Nguyen, B. N. Huynh-Dang, T. B. Nguyen-Tat, UIT_NEWRON at imageclefmed medvqa-gi 2026: Canonical post-processing for gastrointestinal visual question answering, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[34] S. S. H. Emad, I. Safwan, M. A. Tahir, CARES: Calibrated aspect-reliable explanations and safety for gastrointestinal vqa, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[35] P. Netsiri, T. Selenge, F. Kapfer, E. Messeritsch, T. A. Haider, Imageclef medvqa-gi 2026 track 5: Retrieval-augmented generation for medical vqa, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.