Overview of ImageCLEFmedical 2025 – Visual Question Answering and Synthetic Image Generation for 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 third edition of the Medical Visual Question Answering for the Gastrointestinal Tract (MedVQA-GI) challenge, hosted at ImageCLEF 2025. Building on the experiences gained from the last two editions, this year's challenge presented two tasks: (1) Visual Question Answering (VQA) over gastrointestinal (GI) images and (2) high-fidelity synthetic image generation for GI data. Participants were asked to develop multimodal models capable of answering clinical questions based on annotated images and to generate synthetic GI images using text prompts. The dataset was extended from previous years and provided a wide variety of GI images with annotations. Submissions were evaluated using a mix of text generation metrics and image realism metrics. Participation increased slightly from last year, but completion rates remained a challenge. This paper details the tasks, data, evaluation methods, and results. The competition repository is at: github.com/simula/ImageCLEFmed-MEDVQA-GI-2025.

Keywords

Visual question answering, Synthetic medical images, Endoscopy, Machine learning

1. Introduction

The third edition of the Medical Visual Question Answering for the Gastrointestinal Tract (MedVQA-GI) challenge at ImageCLEF continued our focus on advanced image-based machine learning for gastrointestinal (GI) diagnostics. This year we expanded the challenge to include both question answering and text-to-image synthesis tasks. These additions aim to better simulate real-world diagnostic settings by incorporating both image interpretation and generation capabilities into 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]. More recently, there has been growing attention on the use of generative models to create synthetic GI images [24, 25]. These images serve as privacy-preserving alternatives to real data and can be useful for model development, clinician training, and system evaluation. To reflect these trends, this year's MedVQA-GI challenge incorporates both diagnostic reasoning via VQA and synthetic image generation. 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 creation and structure. Then, we present the two challenge tasks along with the evaluation methodology. Finally, we discuss the submissions, results, and lessons learned from this year.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*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)

[1]https://github.com/simula/ImageCLEFmed-MEDVQA-GI-2025

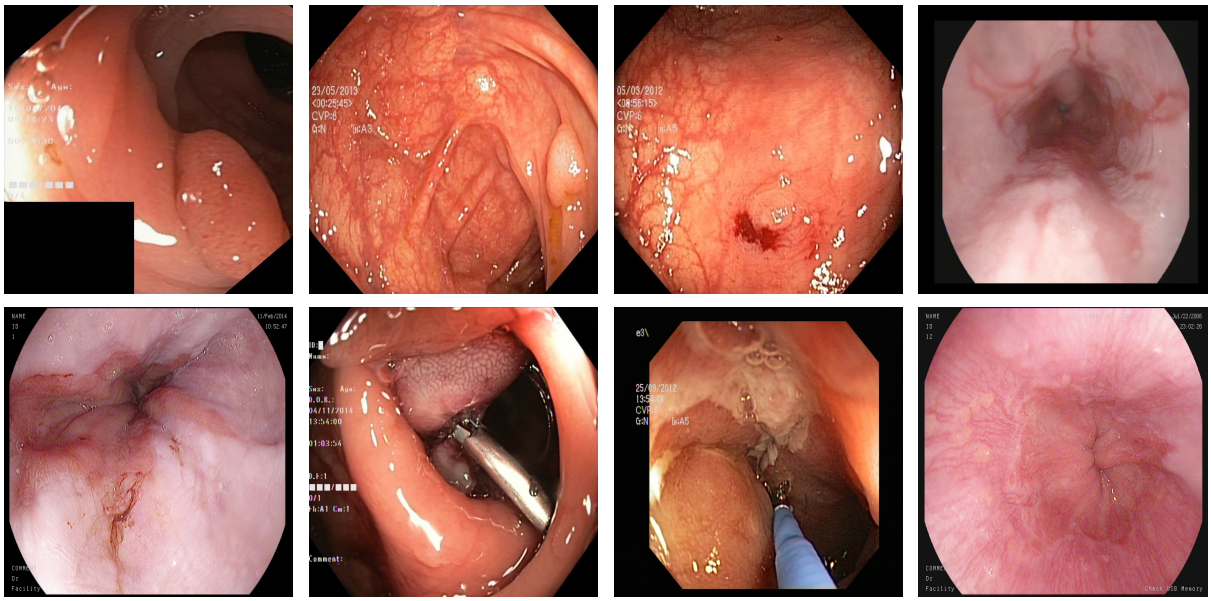

Figure 1: Examples from the development dataset provided by the challenge organizers. The images represent various types of findings and contexts.

2. Dataset

The dataset used in this challenge builds upon the publicly available HyperKvasir [26] and Kvasir-VQA [27] datasets. These datasets consist of gastrointestinal (GI) endoscopy images covering a wide range of anatomical sites and pathological findings, making them suitable for multimodal tasks such as visual question answering (VQA) and image captioning. Examples from the dataset can be seen in Figure 1.

For subtask 1, the development dataset was based on Kvasir-VQA, which contains over 6,500 GI images annotated with visual questions and corresponding answers. The questions span multiple types—Yes/No, Single-Choice, Multiple-Choice, Color, Location, and Count—designed to evaluate a model's capabilities in classification, reasoning, spatial localization, and attribute recognition. Each image was annotated with one or more questions to ensure multimodal diversity and support a range of inference challenges. The dataset reflects clinically relevant scenarios, helping models generalize to real-world diagnostic tasks. The test dataset for subtask 1 was drawn from a custom, unreleased set of GI images. These were sampled from a combination of different sources not included in the development set, ensuring that the test data was distributionally distinct and unseen. This was done to better evaluate the generalization performance of participating systems under realistic conditions.

For subtask 2, participants were provided with a set of over 2,000 image-caption pairs. These were curated to reflect clinically meaningful descriptions of GI endoscopy images, with captions written to summarize findings such as anatomical features, abnormalities, or procedural contexts. To supplement the limited size of the manually annotated caption dataset, a set of additional synthetic captions was released. These were generated using large language models and rule-based methods to provide a diverse range of phrasings and improve the effectiveness of model fine-tuning. The synthetic data aimed to introduce variation and reduce overfitting on the manually annotated samples. As with the VQA task, the captioning test set was drawn from a secret, mixed-source dataset that was distinct from the development data.

3. Tasks and Evaluation

This years, MedVQA-GI is made up of two subtasks: answering clinical questions from GI images and generating synthetic GI images from prompts.

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

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

3.1. Subtask 1: Question Interpretation and Response

This subtask requires participants to submit ML models capable of answering questions based on gastrointestinal (GI) images from the Kvasir-VQA dataset [27]. The dataset consists of $6,500$ annotated images representing a range of anatomical sites, pathological conditions, and endoscopic tools. Each image is paired with a clinical question that falls into one of six categories: Yes/No, Single-Choice, Multiple-Choice, Color-Related, Location-Related, and Numerical Count. These categories require models that can handle both fine-grained visual recognition and contextual understanding of medical language. Questions may require identifying instruments, estimating quantities (like number of polyps), recognizing colors (like bleeding or bile), or locating anatomical features. Model performance is assessed using standard natural language generation metrics:

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.

All models were submitted and validated through a Hugging Face-hosted repository[2]. This setup ensures fair comparison, reproducibility, and allowed participants to view there standing in a publicly registered leaderboard.

3.2. Subtask 2: Synthetic Image Generation

This subtask involves generating synthetic gastrointestinal (GI) images from structured clinical prompts, aiming to mimic the visual and diagnostic complexity of real-world endoscopic imagery. These synthetic outputs are intended to support AI development by enhancing data availability while minimizing dependence on sensitive patient data. Prompts provided to participants detailed anatomical sites, pathological cues, and procedural contexts. The challenge was to synthesize images that closely align with these clinical descriptions while maintaining variability and realism.

To assess model performance, we employed both automatic and expert-driven evaluations. Automated assessment was conducted using four quantitative metrics designed specifically for the medical imaging domain [31]. Each metric was computed using BiomedCLIP [32] image embeddings to ensure clinical relevance:

Fidelity Quantifies visual realism by comparing each generated image to its real counterpart. It is defined as:

$$\text{Fidelity} = \frac{1000}{1 + \text{mean-FID}(A_i, R_i)}$$

where A_i and R_i denote the BiomedCLIP features of generated and real images for prompt i. A higher score reflects closer alignment with real images.

Agreement Measures semantic and visual consistency between images produced from original

[2] https://simulamet-medvqa.hf.space

prompts and their reworded variants. Computed as the mean cosine similarity:

$$\text{Agreement} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|A_i||B_i|}\sum_{a\in A_i, b\in B_i}\frac{a\cdot b}{\|a\|\|b\|}$$

where A_i and B_i are the BiomedCLIP embedding sets of images from the original and rephrased prompts, respectively.

Diversity Captures intra-prompt variability by evaluating the average pairwise Euclidean distance between embeddings of images generated from the same prompt:

$$\text{Diversity} = \frac{1}{M}\sum_{i=1}^{M}\text{pdist}(F_i)$$

with F_i representing the normalized embedding set per prompt and pdist indicating the average pairwise distance function.

Fréchet BiomedCLIP Distance (FBD) Evaluates global distributional alignment between the full sets of synthetic and real images using the Fréchet distance:

$$\text{FBD} = \|\mu_{\text{gen}} - \mu_{\text{real}}\|^2 + \text{Tr}\left(\Sigma_{\text{gen}} + \Sigma_{\text{real}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{real}})^{1/2}\right)$$

where μ and Σ refer to the mean and covariance of BiomedCLIP features. Lower FBD values indicate better overall realism.

In addition to these automated measures, expert raters assessed the clinical plausibility and diagnostic utility of the generated outputs. Like Subtask 1, all submissions were hosted on our public Hugging Face repository.

4. Participation

In total, 45 teams registered for Subtask 1 and 44 for Subtask 2, representing an increase in registrations compared to last year. Of these, 5 teams submitted runs, and 5 teams submitted working notes papers [33, 34, 34, 35, 36, 37]. Table 1 shows an overview of the participants and the number of submissions to each sub-task, alongside the number of participants from last year's challenge. As in previous years, we observed that many who registered did not submit, which is a common pattern. However, we also saw an increase in the number of actual submissions compared to last year. This suggests growing engagement among those who proceed past registration. Future editions could still benefit from improved outreach and support, such as tutorials or "getting started" scripts, to make it easier to participate.

5. Results

Five teams submitted runs to Subtask 1, while three teams participated in Subtask 2. Below, we briefly describe each team's submission and their approach. The overall results are presented in Table 2 for Subtask 1 and Table 3 for Subtask 2.

5.1. Team Sagarmatha Rangers

Team Sagarmatha Rangers [33] participated in Subtask 1 and used Florence-2 as their base model, fine-tuned on the challenge development dataset. They incorporated domain-specific image augmentations like flipping, jitter, and cropping, and embedded location tokens in the prompt to enhance spatial understanding. Training was conducted using LoRA, and evaluation showed that augmentations improved performance.

Table 2
Results for Task 1.

Team	Repo	Set	BLEU	R1	R2	RL	MET
UPS	krissTewari/Florence-2-vqa-final	Public	0.24	0.87	0.11	0.87	0.48
UPS	krissTewari/Florence-2-vqa-final	Private	0.22	0.88	0.11	0.88	0.49
IReL_IIT_BHU	usr256864/BLIP2_kvasir_FT	Public	0.23	0.83	0.10	0.83	0.46
IReL_IIT_BHU	usr256864/BLIP2_kvasir_FT	Private	0.22	0.92	0.11	0.92	0.50
MedPixel	gauravparajuli/florence2_64_r16	Public	0.21	0.87	0.12	0.86	0.48
MedPixel	gauravparajuli/florence2_64_r16	Private	0.18	0.91	0.11	0.90	0.50
CS_Morgan_Lab	sageofai/Florence-2-full-finetune-v3	Public	0.19	0.84	0.10	0.83	0.46
CS_Morgan_Lab	sageofai/Florence-2-full-finetune-v3	Private	0.18	0.90	0.10	0.90	0.49
Sagarmatha_Rangers	laxuu/Florence-2-vqa_final	Public	0.15	0.81	0.10	0.80	0.44
Sagarmatha_Rangers	laxuu/Florence-2-vqa_final	Private	0.16	0.88	0.10	0.88	0.49

Table 3
Results for Task 2.

Team	Repo	Set	Fid.	Agrmt.	Div.	FBD
CS_Morgan_Lab	sageofai/sageofai-lora-kvasir-trained	Private	0.0268	0.7012	0.7017	1539.31
IReL_IIT_BHU	krissTewari/sd-kvasir-imagen-demo	Private	0.2739	0.7390	0.6481	1694.97
MedPixel	gauravparajuli/SD1.5_Prompt2Image	Private	0.2725	0.7329	0.6722	1694.00

5.2. Team CS_Morgan Lab

Team CS_Morgan Lab [34] participated in both Subtask 1 and Subtask 2. For Subtask 1, they fine-tuned a BLIP2-Flan-T5 model with a ViT-G encoder using a causal language modeling approach. Training was done for 3 epochs on an A100 GPU with AdamW and a batch size of 32. They added post-processing to normalize answer outputs. While comparing with zero-shot models like MiniGPT-4 and LLaVA-1.5, their fine-tuned BLIP2 model achieved superior scores across BLEU, ROUGE, and METEOR metrics. For Subtask 2, they used Stable Diffusion v1.5 fine-tuned with LoRA on image-caption pairs from the GI domain. The model was trained using DreamBooth techniques to improve prompt-image alignment.

5.3. Team MedPixel

Team MedPixel [35] participated in both Subtask 1 and Subtask 2. For Subtask 1, they fine-tuned the Florence2-0.3B model on the Kvasir-VQA dataset using LoRA and optimized hyperparameters through Bayesian search with Optuna. Training was conducted on an RTX A4000 GPU, using gradient accumulation to simulate larger batch sizes. Their best model (batch size 64, LoRA rank 16) achieved strong performance, with a METEOR score of 0.48 and ROUGE-L of 0.86 on the private test set. For Subtask 2, they fine-tuned Stable Diffusion v2.1 using LoRA to synthesize GI endoscopy images from structured prompts.

5.4. Team IReL, IIT (BHU)

Team IReL, IIT (BHU) [36] participated in both Subtask 1 and Subtask 2. For Subtask 1, they fine-tuned the Florence2 model on Kvasir-VQA, using image preprocessing to remove specular highlights and black borders. Training was performed on a single H100 GPU using AdamW and fp16 precision. For Subtask 2, they fine-tuned Stable Diffusion v2-1 using LoRA on four L40 GPUs with synthetic prompts and images at 768×768 resolution. They selected v2-1 based on its better trade-off between quality and compute.

5.5. Team UPS

Team UPS [37] participated in Subtask 1 and explored two approaches: a multimodal Chain-of-Thought (CoT) reasoning method and fine-tuning of generative models. The CoT method used Qwen2-VL to generate rationales before predicting answers through two-stage prompting. In contrast, the fine-tuning strategy involved training BLIP2-Flan-T5-XL and Qwen2-VL using cross-entropy loss with LoRA. Models were trained for 10 epochs on a single A100 GPU. The fine-tuned BLIP2 model performed best out of all other configurations, achieving the best scores.

6. Discussion

In Subtask 1, most teams adopted transformer-based multimodal architectures, with Florence2 being most common. Fine-tuning was typically performed using LoRA, together with hyperparameter optimization and input augmentation. In addition, methods incorporating structured prompting or chain-of-thought reasoning showed potential, particularly for complex question categories that required spatial reasoning or numerical inference, such as location-based and counting tasks. In Subtask 2, three teams submitted models based on fine-tuned variants of Stable Diffusion. Similar to Subtask 1, LoRA was the primary fine-tuning strategy. While all models demonstrated high visual fidelity in image generation, the degree of alignment between prompts and outputs varied. Quantitative evaluation using FBD and prompt-image consistency metrics indicated that current methods are still limited in their ability to generate clinically accurate content. In several cases, generated images appeared realistic but failed to reproduce specific anatomical or pathological features described in the input prompts.

Although we received more submissions this year than last, the number of final submissions is still low compared to the number of registered. This suggests that both subtasks still pose technical and resource challenges. Subtask 2 in particular may benefit from additional baseline models, simplified starter code, and clearer guidelines on expected output structure. The evaluation setup, while automated and reproducible, may need to be complemented with more qualitative human review.

7. Conclusion

This paper presented the 2025 edition of the MedVQA-GI challenge, held as part of ImageCLEF. The challenge included two sub-tasks focused on medical VQA and the generation of synthetic gastrointestinal images. For future editions, we aim to refine and expand the task by providing more comprehensive resources to support participants in getting started.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o 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] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. d. Lange, P. Halvorsen, M. A. Riegler, Singan-seg: Synthetic training data generation for medical image segmentation, PLOS ONE 17 (2022) 1–24. URL: https://doi.org/10.1371/journal.pone.0267976. doi:10.1371/journal.pone.0267976.

[25] D. Yoon, H.-J. Kong, B. S. Kim, W. S. Cho, J. C. Lee, M. Cho, M. H. Lim, S. Y. Yang, S. H. Lim, J. Lee, J. H. Song, G. E. Chung, J. M. Choi, H. Y. Kang, J. H. Bae, S. Kim, Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network, Sci Rep 12 (2022) 261.

[26] 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.

[27] 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.

[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] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv (2025). doi:10.48550/arXiv.2505.05573. arXiv:2505.05573.
[32] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, arXiv (2023). doi:10.48550/arXiv.2303.00915. arXiv:2303.00915.
[33] S. Gaihre, A. Thapa, P. Pokhrel, L. Tiwari, Multimodal ai for gastrointestinal diagnostics: Tackling vqa in imageclefmed-medvqa-gi 2025, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[34] E. P. O. Oluwafemi, M. Hoque, E. F. Akor, R. N. Chowdhury, A. Umar, M. M. Rahman, Solving medical data limitations through ai: Multi-modal vision-language learning for gastrointestinal vqa and synthetic training data generation, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[35] G. Parajuli, Querying gi endoscopy images: A vqa approach, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[36] K. Tewari, S. Pal, Bridging vision and language in gi diagnosis: Florence2 for question answering and stable diffusion for image synthesis, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[37] O. Adjali, Towards better gastrointestinal diagnosis: Evaluating vision-language models for gi vqa, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.