Kvasir-VQA: A Text-Image Pair GI Tract Dataset

Sushant Gautam*
SimulaMet
Oslo, Norway

Andrea Storås
SimulaMet
Oslo, Norway

Cise Midoglu
SimulaMet
Oslo, Norway

Steven A. Hicks
SimulaMet
Oslo, Norway

Vajira Thambawita
SimulaMet
Oslo, Norway

Pål Halvorsen*
SimulaMet
Oslo, Norway

Michael A. Riegler*
SimulaMet
Oslo, Norway

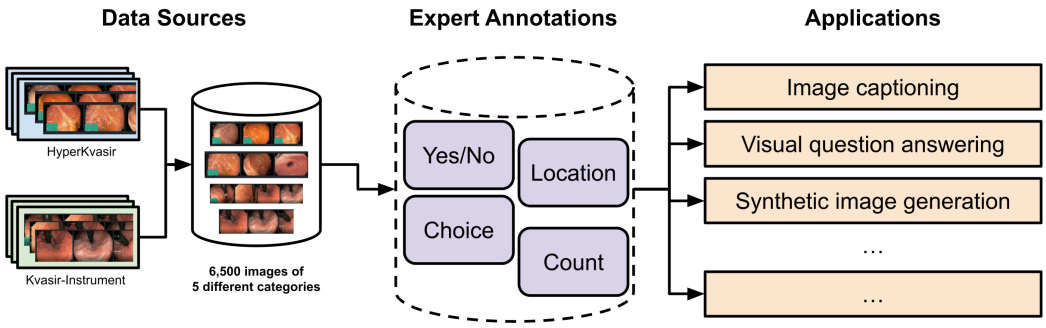


Figure 1: Overview of the sources, curation, and applications of the Kvasir-VQA dataset.

ABSTRACT

We introduce Kvasir-VQA, an extended dataset derived from the HyperKvasir and Kvasir-Instrument datasets, augmented with question-and-answer annotations to facilitate advanced machine learning tasks in Gastrointestinal (GI) diagnostics. This dataset comprises 6,500 annotated images spanning various GI tract conditions and surgical instruments, and it supports multiple question types including yes/no, choice, location, and numerical count. The dataset is intended for applications such as image captioning, Visual Question Answering (VQA), text-based generation of synthetic medical images, object detection, and classification. Our experiments demonstrate the dataset's effectiveness in training models for three selected tasks, showcasing significant applications in medical image analysis and diagnostics. We also present evaluation metrics for each task, highlighting the usability and versatility of our dataset. The dataset and supporting artifacts are available at https://datasets.simula.no/kvasir-vqa.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Machine learning**; **Computer vision tasks**; • **Information systems** → **Question answering**.

KEYWORDS

Medical Image Analysis; Visual Question Answering (VQA); Medical Image Captioning; Gastrointestinal Diagnostics; Machine Learning in Healthcare

1 INTRODUCTION

The advancement of medical diagnostics increasingly relies on the integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques to analyze complex medical images. The human digestive system is categorized into the upper and lower Gastrointestinal (GI) tracts and includes the esophagus, stomach, and intestines. Diseases in the GI tract, being prevalent and often requiring intricate diagnostic procedures, present an ideal domain for deploying AI-driven solutions. Existing datasets such as HyperKvasir [18] and Kvasir-Instrument [27] have made substantial contributions to this field by offering a wide range of classification- and segmentation-labeled GI images.

However, the lack of comprehensive textual annotations, particularly question-and-answer pairs, limits the potential for developing advanced AI models capable of nuanced understanding and decision-making. This limitation is critical because question-and-answer pairs for an input can help to simulate the reasoning process that we humans go through, thereby enabling AI models to better understand context, interpret complex scenarios, and provide more accurate diagnostics. Furthermore, these annotations can facilitate the training of AI systems to handle a wider variety of cases, improving their generalizability and reliability in real-world clinical settings.

In this study, we extend these foundational datasets by incorporating detailed question-and-answer annotations (Figure 1). This enhancement aims to facilitate a broader range of ML applications, such as image captioning, Visual Question Answering (VQA), text-based synthetic image generation, object detection, and classification. By curating a new dataset, Kvasir-VQA, we bridge the gap

*Also affiliated with Oslo Metropolitan University, Oslo, Norway.

between medical image analysis and practical diagnostic tools, ultimately aiming to improve patient outcomes and diagnostic precision. We present preliminary experiments with Kvasir-VQA in order to demonstrate three of these use cases: image captioning, VQA, and synthetic medical image generation. These tasks were selected to highlight different aspects of the dataset and its potential use in various contexts.

The rest of this paper is structured as follows. Section 2 provides a comprehensive background on existing medical image analysis datasets, particularly HyperKvasir and Kvasir-Instrument, and their contributions to the field. Section 3 details our curation of the Kvasir-VQA dataset, including the incorporation of question-and-answer annotations, and provides an elaborate dataset description, specifying the image categories and types of questions included. Section 4 presents the experiments we have conducted to demonstrate the dataset's utility, namely image captioning, VQA, and synthetic medical image generation. In Section 5, we summarize the key findings, discuss the potential applications of our dataset in medical diagnostics as well as its limitations, and outline directions for future research. Section 7 concludes the paper.

2 BACKGROUND AND RELATED WORK

The human gastrointestinal (GI) tract is susceptible to a wide range of abnormal mucosal conditions, varying from minor irritations to highly lethal diseases [22, 43]. According to the International Agency for Research on Cancer, a specialized cancer agency of the World Health Organization (WHO), GI cancers account for approximately 4.8 million new cases annually worldwide [11]. These cancers often have a high mortality rate, contributing to around 3.4 million deaths each year [11].

Endoscopy is the current gold-standard procedure for examining the GI tract, but its effectiveness is limited by the operator's performance, which results in a significant average miss rate of around 20% for polyps in the colon [4, 60]. Consequently, improving endoscopic performance, enhancing the quality of clinical examinations, and implementing systematic screening are crucial for reducing morbidity and mortality associated with GI diseases [31, 51].

The emergence of AI-enabled support systems offers promise in assisting healthcare professionals in providing high-quality care on a large scale [8]. These AI systems, particularly those using ML techniques, require extensive training on well-curated datasets containing human-verified annotations to be effective in real-world tasks, such as detecting precancerous lesions or cancers in medical images [34]. The performance of AI in medical image analysis has seen significant advancements, primarily driven by the quality of the datasets and the sophistication of the algorithms employed [9, 45].

2.1 Gastrointestinal Image Datasets

Earlier GI datasets have focused on different findings such as polyps [10, 12–16, 29, 47, 54, 57], endoscopic artifacts [5], GI lesions [2, 3, 20, 38, 39], angiectasia, bleeding, inflammation, esophagitis, ulcerative colitis, Z/line, pylorus, cecum, dyed resection margins, and stool [10, 12, 14, 35, 46, 47]. However, a number of these datasets are not publicly available, serve more as educational databases rather than being suitable for algorithm training, and/or are not usable for ML. Several other datasets have also focused specifically on diagnostic and therapeutic tool segmentation in endoscopy [6, 7, 17, 27, 49].

Kvasir [48], HyperKvasir [18], Kvasir-SEG [29], Kvasir-Capsule [55] and Kvasir-Instrument [27] are prominent datasets that have catalyzed research in GI diagnostics, see for example [21, 28, 56, 58, 61], providing a large number of GI images of various types and some with bounding boxes and segmentation masks. Recent studies have leveraged these datasets to develop models for specific tasks, such as polyp detection using Convolutional Neural Networks (CNNs) and capsule endoscopy analysis. However, these models are limited by the scope of available annotations, which primarily focus on image classification, object detection, and segmentation.

2.2 Image Captioning

Image captioning in medical imaging has advanced significantly with the introduction of transformer-based models [53]. These models, particularly those leveraging architectures like the Vision Transformer (ViT) and multimodal transformers, have demonstrated superior performance in generating accurate and contextually rich descriptions of medical images [23]. The ability of transformers to capture long-range dependencies and context within images makes them ideal for medical applications where subtle differences can be diagnostically significant [42]. In the context of the Kvasir-VQA dataset, the integration of fine-tuning in captioning models can enhance the models' capability to provide detailed and accurate descriptions that can support clinical decision-making and automated reporting.

2.3 Visual Question Answering

VQA is an emerging research area that combines image understanding with Natural Language Processing (NLP) to answer questions about images [37]. While VQA has seen success in general domains, its application in medical imaging remains nascent due to a lack of specialized datasets [30]. The introduction of question-and-answer annotations in our extended dataset Kvasir-VQA addresses this gap, providing a rich resource for training VQA models tailored to medical diagnostics.

VQA in medical imaging has benefited from transformer-based models, which excel at integrating visual and textual information [41]. Models like BERT and its variants and successors, well-established in NLP, have been adapted for multimodal tasks including VQA [33, 52]. These models effectively understand and generate responses to complex medical questions by synthesizing information from images and text [36]. The Kvasir-VQA dataset, with its rich annotations, provides a critical resource for training these models, enabling the development of AI systems that can deliver accurate and context-aware responses to medical queries, thereby enhancing diagnostic accuracy and efficiency.

2.4 Synthetic Medical Image Generation

Recent advancements in synthetic medical image generation have been significantly driven by diffusion-based models, particularly stable diffusion techniques [32, 44]. These models, a class of generative models, iteratively refine images starting from noise, resulting in highly realistic synthetic outputs [44].

The Kvasir-VQA dataset can be used to train models for text-to-image synthesis to generate synthetic medical images that are not only visually accurate but also reflective of the wide range of conditions seen in real-world clinical settings. These synthetic images can augment real-world data, address issues of data scarcity, and enhance the training of AI systems for rare or complex conditions, thereby improving diagnostic tools and educational resources [26].

3 THE KVASIR-VQA DATASET

The Kvasir-VQA dataset we present in this work is an extension of the publicly available HyperKvasir [18] and Kvasir Instrument [27] datasets. This extended dataset incorporates question-and-answer ground truth data, developed in collaboration with medical experts. It covers the entire GI tract, including both normal and abnormal findings, as well as images of various surgical instruments used in GI procedures, such as colonoscopies and gastroscopies.

3.1 Dataset Sources

The visual components of the Kvasir-VQA dataset are sourced from the HyperKvasir and Kvasir-Instrument datasets. Table 1 presents an overview of the source image categories used in the curation of the dataset. Each of those categories contains images from either normal cases, cases with significant diseases (polyps, esophagitis, and ulcerative colitis), or cases involving medical instruments.

Image Category	Number of Samples	Source Dataset
Normal	2500	HyperKvasir
Polyps	1000	HyperKvasir
Esophagitis	1000	HyperKvasir
Ulcerative Colitis	1000	HyperKvasir
Instrument	1000	Kvasir-Instrument
TOTAL	6500	

Table 1: Findings, number of images and source dataset.

3.2 Annotation Process

Additional question-and-answer ground truth data were collected with input from medical professionals experienced in GI disease diagnostics. Initial annotations were conducted by computer scientists using LabelBox [1], followed by verification by medical experts[1].

For collecting the ground truth, six types of questions were answered for each image in the dataset, namely: Yes/No questions, single-choice questions, multiple-choice questions, color-related questions, location-related questions, and numerical count questions. Table 2 presents an overview of the question types used for the annotations, along with the questions and answer options provided for each type. The annotations cover various GI aspects, including findings (questions 1, 2, 5, 6, 8, 17, 18), abnormalities (questions 9, 12, 15), anatomical landmarks (questions 10, 13, 16), and instruments (questions 11, 14, 19), as well as multimedia aspects, including image artifacts (question 3) and text (question 4).

[1]Due to time constraints, not all samples were validated by medical experts. A more comprehensive version of the dataset, featuring complete expert verification, is planned for release in the near future.

3.3 Final Dataset

The Kvasir-VQA dataset comprises 6,500 images, each annotated with various question-and-answer pairs. This dataset is fully accessible on HuggingFace. A subset of the Kvasir-VQA dataset, comprising 2,000 images and 20,241 captions derived from the annotations for those images, was recently used in a multimedia retrieval challenge [24], which includes two specific tasks:

(1) **Image Synthesis (IS)**: This task involves participants using text-to-image generative models to create a diverse dataset of medical images based on textual prompts. For instance, participants might receive prompts such as "An early-stage colorectal polyp" and are expected to generate an image that accurately represents this description. The development dataset provided includes prompt and image pairs to aid in developing solutions. During the testing phase, participants will be given a list of prompts and must generate one image per prompt to submit to the organizers.
(2) **Optimal Prompt Generation (OPG)**: This task requires participants to generate images using self-created prompts within defined categories. Examples include generating images with a specific number of polyps, a polyp in a designated area of the image, or a polyp of a particular type and size. Other categories may involve creating images without findings in the esophagus or large bowel, or including specific instruments such as biopsy forceps, metal clips, or tubes. Additionally, participants might need to create images featuring anatomical landmarks like the Z-line, pylorus, or cecum. The evaluation will consider the quality of the synthetic images, the complexity of the models and prompts, and the hardware requirements.

In this work, we chose to use the same subset of 2,000 images across the experiments to showcase different applications of the dataset. It is also worth noting that the images in the subset, along with their annotations, were randomly selected in equal proportions from the four different image categories (excluding the "Normal" category) as listed in Table 1.

The existing annotations for the images were then converted to single-sentence captions using a script. The script generated a templatized caption for each of the annotations. The 20,241 textual captions generated from the annotations in the subset are used in both the image captioning and image generation tasks described below.

A large language model (LLM), LLaMA-3 (7B)[3], has been employed to transform these captions into question-and-answer pairs, making them suitable for the VQA task as well. The model was guided by a carefully designed prompt with explicit instructions, and the output was subjected to a retry mechanism and manual review by domain experts to ensure quality and relevance. The process is explained in detail in the following sections.

4 EXPERIMENTS

In this section, we present three preliminary experiments to demonstrate the effectiveness and applicability of the Kvasir-VQA dataset: Image Captioning, Visual Question Answering (VQA), and Synthetic Medical Image Generation. These tasks were selected to

No.	Type	Question	Answer Options
1	Yes/No	Have all polyps been removed?	No / Yes / Not relevant
2		Is this finding easy to detect?	
3		Is there a green/black box artefact?	
4		Is there text?	
5		Does this image contain any finding?	
6	Choice (Single)	What type of polyp is present?	Paris Ip / Paris IIa / Paris Is
7		What type of procedure is the image taken from?	Capsule Endoscopy / Colonoscopy / Gastroscopy
8		What is the size of the polyp?	<5mm / 5-10mm / 11-20mm / >20mm
9	Choice (Multiple)	Are there any abnormalities in the image?	Oesophagitis, Ulcerative Colitis, Short-Segment Barretts, Barretts, Polyp, Hemorrhoids
10		Are there any anatomical landmarks in the image?	Ileum, Z-Line, Cecum, Pylorus
11		Are there any instruments in the image?	Tube, Metal Clip, Polyp Snare, Injection Needle, Biopsy Forceps
12	Choice (Color)	What color is the abnormality?	landmark:grey, flesh, pink, black, orange, etc.
13		What color is the anatomical landmark?	pink, red, etc.
14	Location	Where in the image is the instrument?	Upper-Left / Upper-Center / Upper-Right / Center-Left / Center / Center-Right / Lower-Left / Lower-Center / Lower-Right
15		Where in the image is the abnormality?	
16		Where in the image is the anatomical landmark?	
17	Numerical Count	How many findings are present?	[0-inf]
18		How many polyps are in the image?	
19		How many instruments are in the image?	

Table 2: Question types with the question and answer options used for the annotations.

highlight different aspects of the dataset and its potential use in various medical imaging and diagnostic contexts. Our implementation for all experiments is available open source.

The results from our experiments underscore the versatility and robustness of the Kvasir-VQA dataset, providing a foundation for future advancements in medical image analysis and diagnostics. The results indicate that models trained on Kvasir-VQA can effectively perform various vision-language tasks, showcasing the dataset's potential in real-world medical applications.

4.1 Image Captioning

Model and Setup: For the image captioning task, we used Florence-2 [62], an open-source, lightweight vision-language model, known for its strong zero-shot and fine-tuning capabilities across various tasks, including captioning, object detection, grounding, and segmentation. We fine-tuned Florence-2 with the prefix <DETAILED_CAPTION>. The model consists of 0.23 billion parameters, and we applied Low-Rank Adaptation (LoRA) [25] for efficient fine-tuning.

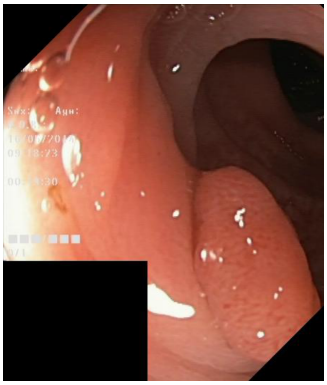
To optimize training efficiency and resource usage, we froze the image encoder during fine-tuning. This decision was based on the observation that while unfreezing the image encoder could potentially enhance performance, it would also significantly increase computational demands and resource consumption. By maintaining the image encoder in a frozen state, we focused the training process on the language model component, allowing us to achieve substantial improvements in caption generation with reduced resource requirements.

During training, we used a learning rate of 1.8e-6 and employed a linear scheduler to manage the learning rate over the course of training. The model was trained for ten epochs with a batch size of 20 on NVIDIA A100 GPU, utilizing the AdamW optimizer to update model weights. We monitored training and validation loss to assess model performance and adjust training parameters as needed. The model and processor were saved at each epoch, facilitating incremental evaluation and potential model deployment.

Dataset and Training: A subset of Kvasir-VQA was employed for this task, where each medical image is paired with a descriptive caption. The whole of the data-subset with 2,000 images was used for the training. The model was trained for 10 epochs, ensuring the loss had stabilized. During inference, the model generates the top 5 candidate captions for each image, acknowledging that different valid aspects of the image might be highlighted.

Evaluation Metrics: We evaluated the fine-tuned model using standard captioning metrics such as Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), and Consensus-based Image Description Evaluation (CIDEr) [59] over 5,000 image-caption pairs from the data-subset. The results are presented in Table 3, demonstrating the model's capability to accurately describe medical images. An example of caption generation[2] from the trained model is presented in Figure 2.

[2]The results are generated by an AI model and may not always accurately reflect true medical conditions or diagnoses.



Captions generated by the model (5 generations):

(1) a polyp of type paris iia
(2) a polyp of type with the color red
(3) a polyp of type paris
(4) a green/black box artefact
(5) a polyp image from a colonoscopy

Figure 2: [Task 1] An example from the fine-tuned captioning model, generating five captions for the given input image.

Task	Metric	Score
Image Captioning	BLEU	0.0823
	ROUGE	0.3905
	METEOR	0.2632
	CIDEr	0.2642
Visual Question Answering	BLEU	0.3757
	ROUGE	0.6955
	METEOR	0.6640
	CIDEr	0.7320
Synthetic Medical Image Generation	FID	110.73
	IS	3.07; 0.14
	IS (real)	4.05; 0.24

Table 3: Model evaluation metrics for the three tasks. FID is Fréchet Inception Distance, and IS is Inception Score. IS (real) represents the Inception Score for the real image dataset used in training. The average and standard deviation of the Inception Score are reported.

4.2 Visual Question Answering

Synthetic VQA Dataset Generation We generated a synthetic VQA dataset by leveraging the capabilities of the LLaMA-3 (7B)[3] language model. The primary objective was to create diverse and contextually relevant question-and-answer pairs from existing image captions, thereby producing a valuable resource for training and evaluating VQA systems.

We commenced with the curated set of image captions. These captions already provided descriptive information about the images, encompassing visible properties, procedural details, and notable findings.

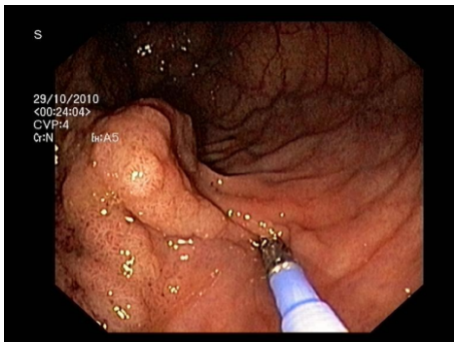
[3]https://github.com/meta-llama/llama3

- Language Model Selection: The LLaMA-3 (7B) model, a state-of-the-art language model known for its adeptness in understanding and generating human-like text, was selected for this task. The model's large parameter size and advanced training regimen equipped it to handle the complex and domain-specific language inherent in medical captions.
- Prompt Design: To guide the language model in generating question-and-answer pairs, we designed a comprehensive prompt, which included:
  - A role specification for the model, indicating its function as an intelligent dataset generator.
  - Explicit instructions to formulate questions based on the content of the captions, ensuring that the questions did not refer directly to the images. This was crucial for maintaining generalizability and focusing on the textual information.
  - Specific guidelines for generating questions related to aspects such as the presence of text, the type of procedure, polyp count, color, size, and location, when mentioned in the caption.
- Data Generation Process: Utilizing the LLaMA-3 (7B) model, for each caption, a JSON object was generated containing the question (Q) and answer (A) pairs. This process involved:
  - Parsing the caption to identify key elements.
  - Crafting a question that emphasizes these elements.
  - Formulating an answer that accurately reflects the information conveyed in the caption.
- Quality Control: To ensure the quality and relevance of the generated data, several quality control measures were implemented:
  - Retry Mechanism: The data generation process included a retry mechanism to handle instances where the output was incomplete or non-informative, thereby ensuring the robustness of the dataset.
  - Manual Review: A subset of the generated question-and-answer pairs was subjected to manual review by domain experts to evaluate the appropriateness of the questions and the correctness of the answers.
- Output and Storage: The resulting question-and-answer pairs were saved, organized according to the index of the original captions. This systematic organization facilitates easy access and integration into VQA systems for training and evaluation.

Model and Setup: To address the limitations of image captioning in controlling specific aspects of image information, we leveraged the VQA task. We fine-tuned the same Florence-2 model with the prefix `<MedVQA>` and included both the input image and the corresponding question in the input. This task benefits from easier evaluation compared to captioning, as the model's output is constrained to answering specific questions about the image.

Dataset and Training: We used the synthetic VQA dataset generated by prompting the LLaMA-3 (7B) language model[3] to create relevant question-and-answer pairs from the existing captions. Similar to the image captioning task, the subset of VQA dataset with 2,000 images was used in the training for synthetic VQA generation,

and 20,241 question-and-answer pairs were synthesized from then captions. The model was fine-tuned for 10 epochs.



(1) **Q:** How many polyps are seen?
A: 1 polyp is seen.
(2) **Q:** What is the color?
A: It's pink.
(3) **Q:** Are there any medical instruments?
A: Yes, there is one.

Figure 3: [Task 2] An example from the fine-tuned VQA model answering questions about the input image.

Evaluation Metrics: We assessed the performance of the VQA model using metrics such as BLEU, ROUGE, METEOR, and CIDEr on all the question-and-answer pairs generated from the subset. The detailed results are shown in Table 3, highlighting the model's proficiency in answering specific medical questions based on image content. An example[2] of question answering from the trained model is presented in Figure 3.

4.3 Synthetic Medical Image Generation

Model and Setup: To generate high-quality synthetic medical images, we trained the Stable Diffusion 3 model [19], which utilizes an advanced Multimodal Diffusion Transformer (MMDiT) architecture. This model translates textual prompts into high-resolution images, thereby enhancing the dataset's utility for training and testing various models. The training employed the DreamBooth technique [50], which updates the entire diffusion model by training on just a few images of a subject or style, associating a special word in the prompt with the example images. A resolution of 512x512 pixels was used, and a large batch size of 48 was employed with gradient accumulation steps set to 1 on the NVIDIA A100 GPU. Mixed precision (fp16) training and the Prodigy optimizer were utilized [40], alongside a sigma-sqrt weighting scheme and a constant learning rate. The model underwent validation at every epoch to evaluate the generated output with set prompts, and gradient checkpointing was enabled to optimize memory usage.

Dataset and Training: The input prompts for the synthetic medical image generation task were derived from the captions used in the image captioning task. We focused on generating high-quality synthetic images that accurately represent the textual descriptions. The model's fine-tuning process included the application of a low-rank adaptation (LoRA) with a rank value of 128, allowing for efficient parameter updates without overfitting. The model, containing 2 billion parameters, was fine-tuned using LoRA for 80 epochs until the training loss had stabilized. Post-training evaluation demonstrated that the model could consistently produce anatomically plausible and diagnostically relevant images, making it a valuable tool for generating medical image datasets.

Generation Prompt: a non polyp image with an instrument

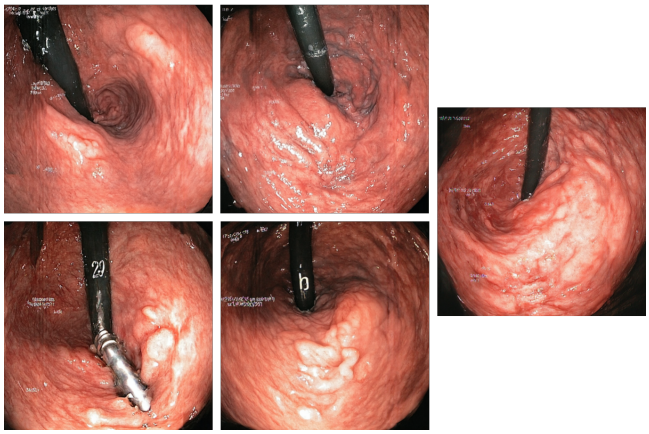


Figure 4: [Task 3] An example from the fine-tuned synthetic medical image generation model, which generated five images for the given prompt.

Evaluation Metrics: We first generated 5,000 images each for the polyp and non-polyp classes using randomly selected prompts from the data subset. The generated images in these two classes were then evaluated using metrics such as Fréchet Inception Distance (FID) and Inception Score (IS) to assess the quality and diversity of the synthetic images. The quality and diversity of the real images in these two classes from the subset were also calculated for comparison using the IS score. Table 3 presents the evaluation results, demonstrating the effectiveness of the model in generating realistic medical images. The FID score and IS for the generated images indicate that while the model produces images with some level of realism, there is still a gap compared to real images. A lower FID score and a higher IS value generally correspond to more realistic and diverse images. Therefore, the results suggest that the generated images, while somewhat effective, still have room for improvement in terms of visual quality and diversity. An example of image generation[2] from the trained model is presented in Figure 4.

5 DISCUSSION

The Kvasir-VQA dataset aims to bridge the gap between medical image analysis and practical diagnostic applications, ultimately contributing to improved patient care and diagnostic accuracy. Below, we elaborate on dataset applications, as well as limitations and potential future work.

5.1 Dataset Applications

The Kvasir-VQA dataset extends the capabilities of existing GI datasets by introducing comprehensive question-and-answer annotations for medical images. This enhancement allows for the

development of sophisticated ML models that can perform a variety of tasks essential for medical diagnostics.

- **Image Captioning**: As demonstrated in Section 4, Kvasir-VQA can be used to train captioning models. By generating descriptive captions for medical images, it is possible to automate the creation of detailed medical reports, reducing the burden on healthcare professionals and minimizing the risk of human error. The Florence-2 model, fine-tuned for the image captioning task, has shown impressive performance in generating accurate and relevant captions, as evidenced by high BLEU, METEOR, and CIDEr scores.
- **Visual Question Answering (VQA))**: The incorporation of VQA tasks enables models to interpret medical images in a more interactive and detailed manner. Our experiments with the Florence-2 model, fine-tuned for VQA, demonstrate the model's ability to accurately answer specific medical questions based on visual input. This capability is crucial for developing diagnostic tools that can assist clinicians by providing immediate, context-specific information about GI conditions. The ability for the model to provide text as output might also enhance healthcare professionals' interaction with the model because the prediction includes more context than just the predicted class or diagnosis. This can make it easier to 'quality check' the model´s reasoning behind the prediction and contribute to a more natural way of interacting with a machine.
- **Synthetic Medical Image Generation**: The ability to generate high-quality synthetic medical images using models such as Stable Diffusion 3 opens new avenues for data augmentation and training. Synthetic images can help address the issue of class imbalance in medical datasets and provide additional training samples for rare conditions. Our experiments demonstrate the effectiveness of Kvasir-VQA for synthetic medical image generation, with generated images exhibiting high fidelity and diversity as measured by FID and IS metrics.
- **Object Detection and Localization**: The question-and-answer annotations related to the location of abnormalities, instruments, and anatomical landmarks enable precise training of object detection and localization models. These models are essential for tasks such as polyp detection and surgical instrument recognition, which are critical for real-time diagnostic support during endoscopic procedures.
- **Classification**: The yes/no and choice questions in the Kvasir-VQA dataset facilitate the development of classification models. These models can assist in diagnosing specific GI conditions, identifying procedural contexts, and recognizing the presence of surgical instruments. Kvasir-VQA supports both single-choice and multiple-choice questions, allowing for a comprehensive evaluation of classification model performance.

5.2 Synthetic VQA Dataset Generation

The synthetic VQA dataset generated with LLM demonstrated a broad vocabulary and diverse question structures, showcasing the model's ability to understand and interpret complex medical language. This linguistic variety is crucial for developing robust VQA systems that can handle a wide range of questions and scenarios in medical imaging.

Using the LLaMA-3 (7B) model to generate synthetic VQA datasets has proven effective, offering a scalable and efficient method for producing large datasets with varied linguistic patterns. This approach not only improves training data for VQA systems but also advances medical AI by fostering the development of more nuanced models. Future work could refine prompt engineering and integrate iterative feedback to further enhance dataset quality.

6 LIMITATIONS AND FUTURE WORK

While the Kvasir-VQA dataset marks significant progress, it is crucial to acknowledge that several limitations remain. Addressing these limitations presents opportunities for meaningful improvements and further advancements in the dataset's utility and accuracy.

- **Expert Verification of Annotations:** As outlined in Section 3, the initial annotations in the Kvasir-VQA dataset were performed by computer scientists, with subsequent verification by medical professionals. However, due to time constraints, not all annotations received expert validation. Future work will include the release of an enhanced version of the dataset, which will undergo complete expert verification to ensure the highest standards of annotation accuracy and reliability.
- **Scope:** The current Kvasir-VQA dataset includes samples from a range of GI conditions and procedural contexts. However, it does not cover the full spectrum of GI conditions encountered in clinical practice. Future efforts will focus on expanding the dataset to encompass a wider array of GI conditions and procedural contexts, thereby increasing its comprehensiveness and applicability for diverse diagnostic applications.
- **Scale:** The Kvasir-VQA dataset currently comprises 6,500 annotated images, a subset of the images available in the source datasets, such as HyperKvasir, which contains over 10,662 labeled images.
  We aim to expand the dataset by integrating additional images and annotations from these source datasets, thereby providing a more extensive resource for training and evaluating machine learning models.
- **Validation:** All experiments and results reported in this study are based on a subset of 2,000 images, including their corresponding annotations, derived captions, and question-and-answer pairs. For reproducibility, comparative analysis, and proper validation of the trained models, the full dataset should be utilized with appropriate training, validation, and test data splits. Addressing this will be considered in future work.
- **Development of Diagnostic Tools:** While the Kvasir-VQA dataset is designed to support the development of AI-driven diagnostic tools, the current iteration may not fully address all necessary components for comprehensive diagnostic support. Future enhancements will aim to align the

dataset with emerging clinical requirements and technological advancements, thereby more effectively contributing to the development and refinement of diagnostic tools and systems in healthcare.

Future work will prioritize the augmentation of the dataset through additional expert-verified annotations and an expanded scope that includes a broader range of GI conditions and procedural contexts. By continuously improving and broadening this resource, we aspire to facilitate the development of advanced AI-driven diagnostic tools that have the potential to transform medical imaging and diagnostics.

7 CONCLUSION

In this paper, we introduce the Kvasir-VQA dataset, an extension of the existing HyperKvasir and Kvasir-Instrument datasets, enriched with question-and-answer annotations. This dataset is specifically designed to advance research in medical image analysis, particularly in the field of gastrointestinal diagnostics. The Kvasir-VQA dataset facilitates a variety of applications, including image captioning, VQA, and synthetic medical image generation, as demonstrated through our preliminary experiments.

The introduction of synthetic question-and-answer pairs provides a new dimension to the dataset, enabling the development of sophisticated AI models capable of nuanced understanding and interactive diagnostics. These capabilities are crucial for improving the accuracy and efficiency of medical diagnostics, thereby improving patient care. Despite its promising potential, the Kvasir-VQA dataset is still a work in progress. Future efforts will focus on extending the dataset's scope and scale, ensuring comprehensive expert validation of annotations, and aligning with clinical needs. These enhancements will further solidify the dataset's role as a valuable resource for the medical community, supporting the development of next-generation diagnostic tools.

ACKNOWLEDGMENTS

This work has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

REFERENCES

[1] 2024. Labelbox | Data factory for the next GenAI. https://labelbox.com [Online; accessed 30. Jul. 2024].

[2] 2024. The Gastrolab Image Gallery. http://www.gastrolab.net/index.htm [Online; accessed 30. Jul. 2024].

[3] 2024. WEO Endoscopy Atlas: Search the Atlas. http://www.endoatlas.org/index.php [Online; accessed 30. Jul. 2024].

[4] Omer F. Ahmad, Antonio S. Soares, Evangelos Mazomenos, Patrick Brandao, Roser Vega, Edward Seward, Danail Stoyanov, Manish Chand, and Laurence B. Lovat. 2019. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *Lancet Gastroenterology & Hepatology* 4, 1 (Jan. 2019), 71–80. https://doi.org/10.1016/S2468-1253(18)30282-6

[5] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnières, Victor Loschenov, Enrico Grisan, Walter Blondel, and Jens Rittscher. 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. *arXiv preprint arXiv:1905.03209* (2019).

[6] Max Allan et al. 2019. 2017 Robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426* (2019).

[7] Max Allan and Mahdi Azizian. 2019. Robotic Scene Segmentation Sub-Challenge. arXiv preprint arXiv:1902.06426.

[8] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med. Educ.* 23, 1 (Dec. 2023), 1–15. https://doi.org/10.1186/s12909-023-04698-z

[9] Fouzia Altaf, Syed M. S. Islam, Naveed Akhtar, and Naeem Khalid Janjua. 2019. Going Deep in Medical Image Analysis: Concepts, Methods, Challenges, and Future Directions. *IEEE Access* 7 (July 2019), 99540–99572. https://doi.org/10.1109/ACCESS.2019.2929365

[10] Quentin Angermann, Jorge Bernal, Cristina Sánchez-Montes, Maroua Hammami, Gloria Fernández-Esparrach, Xavier Dray, Olivier Romain, F Javier Sánchez, and Aymeric Histace. 2017. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE CLIP)*. Vol. 10550. 29–41. https://doi.org/10.1007/978-3-319-67543-5_3

[11] Melina Arnold, Christian C. Abnet, Rachel E. Neale, Jerome Vignat, Edward L. Giovannucci, Katherine A. McGlynn, and Freddie Bray. 2020. Global Burden of 5 Major Types of Gastrointestinal Cancer. *Gastroenterology* 159, 1 (July 2020), 335–349.e15. https://doi.org/10.1053/j.gastro.2020.02.068

[12] Jorge Bernal and Histace Aymeric. 2017. Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D&L challenge. https://endovissub2017-giana.grand-challenge.org/home/. Accessed: 2017-11-20.

[13] Jorge Bernal and Histace Aymeric. 2017. MICCAI Endoscopic Vision Challenge Polyp detection and segmentation. https://endovissub2017-giana.grand-challenge.org/home/. Accessed: 2017-12-11.

[14] Jorge Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodriguez, Maroua Hammami, Ana Garcia-Rodriguez, Henry Córdova, Olivier Romain, Gloria Fernández-Esparrach, Xavier Dray, and Javier Sanchez. 2018. 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)*. https://doi.org/hal-01846141

[15] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* 43 (2015), 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007

[16] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 9 (2012), 3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002

[17] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475* (2018).

[18] Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K. Stensland, Enrique Garcia-Ceja, Peter T. Schmidt, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, and Thomas de Lange. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* 7, 283 (Aug. 2020), 1–14. https://doi.org/10.1038/s41597-020-00622-y

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv* (March 2024). https://doi.org/10.48550/arXiv.2403.03206 arXiv:2403.03206

[20] Julio Murra-Saca et al. 2019. El Salvador Atlas of Gastrointestinal Video Endoscopy. http://www.gastrointestinalatlas.com/index.html. Accessed: 2019-12-16.

[21] Jan Andre Fagereng, Vajira Thambawita, Andrea M. Storås, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A. Riegler. 2022. PolypConnect: Image inpainting for generating realistic gastrointestinal tract images with polyps. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. 66–71. https://doi.org/10.1109/CBMS55023.2022.00019

[22] H. Gelberg. 2018. Pathophysiological Mechanisms of Gastrointestinal Toxicity. *Comprehensive Toxicology* (2018), 139. https://doi.org/10.1016/B978-0-12-801238-3.10923-7

[23] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. 2023. Transformers in medical image analysis. *Intelligent Medicine* 3, 1 (Feb. 2023), 59–78. https://doi.org/10.1016/j.imed.2022.07.002

[24] Steven Hicks, Andrea M Storås, Pål Halvorsen, Thomas de Lange, Michael Riegler, and Vajira Thambawita. 2023. Overview of ImageCLEFmedical 2023-Medical Visual Question Answering for Gastrointestinal Tract. In *CLEF (Working Notes)*. 1316–1327.

[25] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large

Language Models. *arXiv* (June 2021). https://doi.org/10.48550/arXiv.2106.09685 arXiv:2106.09685

[26] Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gokhan Ertaylan, and Michel Dumontier. 2024. Generative AI for Synthetic Data Across Multiple Medical Modalities: A Systematic Review of Recent Developments and Challenges. *arXiv* (June 2024). https://doi.org/10.48550/arXiv.2407.00116 arXiv:2407.00116

[27] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A. Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A. Riegler, Thomas de Lange, Peter T. Schmidt, Håvard D. Johansen, Dag Johansen, and Pål Halvorsen. 2021. Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy. In *MultiMedia Modeling*. Springer, Cham, Switzerland, 218–229. https://doi.org/10.1007/978-3-030-67835-7_19

[28] Debesh Jha, Sharib Ali, Steven Hicks, Vajira Thambawita, Hanna Borgli, Pia H. Smedsrud, Thomas de Lange, Konstantin Pogorelov, Xiaowei Wang, Philipp Harzig, Minh-Triet Tran, Wenhua Meng, Trung-Hieu Hoang, Danielle Dias, Tobey H. Ko, Taruna Agrawal, Olga Ostroukhova, Zeshan Khan, Muhammad Atif Tahir, Yang Liu, Yuan Chang, Mathias Kirkerød, Dag Johansen, Mathias Lux, Håvard D. Johansen, Michael A. Riegler, and Pål Halvorsen. 2021. A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging. *Medical Image Analysis* 70 (2021), 102007. https://doi.org/10.1016/j.media.2021.102007

[29] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *Proceeding of International Conference on Multimedia Modeling (MMM)*, Vol. 11962. 451–462. https://doi.org/10.1007/978-3-030-37734-2_37

[30] Vasudha Joshi, Pabitra Mitra, and Supratik Bose. 2024. Multi-modal multi-head self-attention for medical VQA. *Multimed. Tools Appl.* 83, 14 (April 2024), 42585–42608. https://doi.org/10.1007/s11042-023-17162-3

[31] Michal F. Kaminski, Siwan Thomas-Gibson, Marek Bugajski, Michael Bretthauer, Colin J. Rees, Evelien Dekker, Geir Hoff, Rodrigo Jover, Stepan Suchanek, Monika Ferlitsch, John Anderson, Thomas Roesch, Rolf Hultcranz, Istvan Racz, Ernst J. Kuipers, Kjetil Garborg, James E. East, Maciej Rupinski, Birgitte Seip, Cathy Bennett, Carlo Senore, Silvia Minozzi, Raf Bisschops, Dirk Domagk, Roland Valori, Cristiano Spada, Cesare Hassan, Mario Dinis-Ribeiro, and Matthew D. Rutter. 2017. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy* 49, 04 (April 2017), 378–397. https://doi.org/10.1055/s-0043-103411

[32] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2023. Diffusion models in medical imaging: A comprehensive survey. *Med. Image Anal.* 88 (Aug. 2023), 102846. https://doi.org/10.1016/j.media.2023.102846

[33] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U. Deva Priyakumar, and C. V. Jawahar. 2021. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 13–16. https://doi.org/10.1109/ISBI48211.2021.9434063

[34] Dow-Mu Koh, Nickolas Papanikolaou, Ulrich Bick, Rowland Illing, Charles E. Kahn, Jayshree Kalpathi-Cramer, Celso Matos, Luis Martí-Bonmatí, Anne Miles, Seong Ki Mun, Sandy Napel, Andrea Rockall, Evis Sala, Nicola Strickland, and Fred Prior. 2022. Artificial intelligence and machine learning in cancer imaging. *Commun. Med.* 2, 133 (Oct. 2022), 1–14. https://doi.org/10.1038/s43856-022-00199-0

[35] Anastasios Koulaouzidis, Dimitris K. Iakovidis, Diana E. Yung, Emanuele Rondonotti, Uri Kopylov, John N. Plevris, Ervin Toth, Abraham Eliakim, Gabrielle Wurm Johansson, Wojciech Marlicz, Georgios Mavrogenis, Artur Nemeth, Henrik Thorlacius, and Gian Eugenio Tontini. 2017. KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open* 5, 6 (May 2017), E477–E483. https://doi.org/10.1055/s-0043-105488

[36] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artif. Intell. Med.* 143 (Sept. 2023), 102611. https://doi.org/10.1016/j.artmed.2023.102611

[37] Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Multiscale Feature Extraction and Fusion of Image and Text in VQA. *Int. J. Comput. Intell. Syst.* 16, 1 (Dec. 2023), 1–11. https://doi.org/10.1007/s44196-023-00233-6

[38] David M. Martin. 2019. The Atlas of Gastrointestinal Endoscope. http://www.endoatlas.com/atlas_1.html. Accessed: 2019-12-12.

[39] Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. 2019. Gastrointestinal Lesions in Regular Colonoscopy Dataset. http://www.depeca.uah.es/colonoscopy_dataset/. Accessed: 2019-12-12.

[40] Konstantin Mishchenko and Aaron Defazio. 2023. Prodigy: An Expeditiously Adaptive Parameter-Free Learner. *arXiv* (June 2023). https://doi.org/10.48550/arXiv.2306.06101 arXiv:2306.06101

[41] Usman Naseem, Matloob Khushi, and Jinman Kim. 2022. Vision-Language Transformer for Interpretable Pathology Visual Question Answering. *IEEE J. Biomed. Health Inf.* 27, 4 (March 2022), 1681–1690. https://doi.org/10.1109/JBHI.2022.3163751

[42] Khalid Nassiri and Moulay A. Akhloufi. 2024. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics* 4, 2 (April 2024), 1097–1143. https://doi.org/10.3390/biomedinformatics4020062

[43] Christine B. Navarre and D. G. Pugh. 2002. Diseases of the Gastrointestinal System. *Sheep & Goat Medicine* (2002), 69. https://doi.org/10.1016/B0-72-169052-1/50006-5

[44] Shaoyan Pan, Tonghe Wang, Richard L. J. Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B. Patel, Joseph Shelton, Sagar A. Patel, Justin Roper, and Xiaofeng Yang. 2023. 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. *Phys. Med. Biol.* 68, 10 (May 2023), 105004. https://doi.org/10.1088/1361-6560/acca5c

[45] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. 2020. AI in Medical Imaging Informatics: Current Challenges and Future Directions. *IEEE J. Biomed. Health Inf.* 24, 7 (May 2020), 1837–1857. https://doi.org/10.1109/JBHI.2020.2991043

[46] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*. 170–174. https://doi.org/10.1145/3083187.3083216

[47] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the ACM Multimedia Systems Conference (ACM MMSYS)*. 164–169. https://doi.org/10.1145/3083187.3083212

[48] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, 164–169. https://doi.org/10.1145/3083187.3083212

[49] Tobias Ross et al. 2020. Robust Medical Instrument Segmentation Challenge 2019. *arXiv preprint arXiv:2003.10299* (2020).

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. [n. d.]. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 17–24. https://doi.org/10.1109/CVPR52729.2023.02155

[51] Adrian Săftoiu, Cesare Hassan, Miguel Areia, Manoop S. Bhutani, Raf Bisschops, Erwan Bories, Irina M. Cazacu, Evelien Dekker, Pierre H. Deprez, Stephen P. Pereira, Carlo Senore, Riccardo Capocaccia, Giulio Antonelli, Jeanin van Hooft, Helmut Messmann, Peter D. Siersema, Mario Dinis-Ribeiro, and Thierry Ponchon. 2020. Role of gastrointestinal endoscopy in the screening of digestive tract cancers in Europe: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 52, 04 (April 2020), 293–304. https://doi.org/10.1055/a-1104-5245

[52] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K. Krishna, and Hongliang Ren. 2022. Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer, Cham, Switzerland, 33–43. https://doi.org/10.1007/978-3-031-16449-1_4

[53] Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V. Dylov. 2023. Medical image captioning via generative pretrained transformers. *Sci. Rep.* 13, 4171 (March 2023), 1–12. https://doi.org/10.1038/s41598-023-31223-5

[54] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery* 9, 2 (2014), 283–293. https://doi.org/10.1007/s11548-013-0926-3

[55] Pia H. Smedsrud, Vajira Thambawita, Steven A. Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc Tien Dang Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T. Schmidt, Ervin Toth, Hugo L. Hammer, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. 2021. Kvasir-Capsule, a video capsule endoscopy dataset. *Sci. Data* 8, 142 (May 2021), 1–10. https://doi.org/10.1038/s41597-021-00920-z

[56] Abhishek Srivastava, Nikhil Kumar Tomar, Ulas Bagci, and Debesh Jha. 2022. Video Capsule Endoscopy Classification using Focal Modulation Guided Convolutional Neural Network. In *IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 21–23. https://doi.org/10.1109/CBMS55023.2022.00064

[57] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. 2016. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging* 35, 2 (2016), 630–644. https://doi.org/10.1109/TMI.2015.2487997

[58] Nefeli Panagiota Tzavara and Bjørn-Jostein Singstad. 2021. Transfer learning in polyp and endoscopic tool segmentation from colonoscopy images. *Nordic Machine Intelligence* 1, 1 (2021), 32–34. https://doi.org/10.5617/nmi.9132

[59] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. [n. d.]. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7299087

[60] Pu Wang, Tyler M. Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. 2019. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 68, 10 (Oct. 2019), 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

[61] Xing Wu, Cheng Chen, Mingyu Zhong, and Jianjia Wang. 2021. HAL: Hybrid active learning for efficient labeling in medical domain. *Neurocomputing* 456 (2021), 563–572. https://doi.org/10.1016/j.neucom.2020.10.115

[62] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2023. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. *arXiv* (Nov. 2023). https://doi.org/10.48550/arXiv.2311.06242 arXiv:2311.06242