Generative AI Based on Medical Visual Question Answering (VQA) Techniques

**by**  
**[Sarthak Kaushik]**

Thesis Submitted in Partial Fulfillment of the  
Requirements for the Degree of  
Master of Computer Science

in the  
[Department of Computer Science]  
[Faculty of Computer Science]

© [Sarthak Kaushik] 2025  
LAKEHEAD UNIVERSITY  
[April 2026]

Copyright in this work is held by the author. Please ensure that any reproduction   
or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

|  |  |  |
| --- | --- | --- |
| **Name:** | **Sarthak Kaushik** | |
| **Degree:** | **Master of Computer Science** | |
| **Title:** | **Generative AI Based on Medical Visual Question Answering (VQA) Techniques** | |
| **Committee:** | **Chair:** | **[Dr. Sabah Mohammed]**  Supervisor [Chair, Department of Computer Science] |
|  | **[Dr. Jinan Fiaidhi]** Supervisor [Full Professor, Department of Computer Science] | |
|  | **[Dr. Arnold Kim]** External Examiner  [Academic or Professional Role, Academic Unit] [Institution] | |

Ethics Statement

Abstract

Visual Question Answering (VQA) allows an artificial-intelligence agent to interpret an image and respond to a clinician’s question in natural language. This thesis develops the first transformer-based MedVQA framework devoted to gastrointestinal endoscopy, with a focus on colonoscopy images. Leveraging the new Kvasir-VQA and Kvasir-VQA-x1 datasets, the system couples a domain-adapted vision encoder with biomedical language models (BioBERT for question understanding and BioGPT for answer generation) to handle both closed-form and free-text questions. Rigorous evaluation on standard and perturbed images shows that the model answers lesion-level queries, such as polyp count and location, with accuracy comparable to state-of-the-art detectors while adding human-readable explanations. A dedicated module grades ulcerative colitis severity by answering questions framed in Mayo or UCEIS terms, matching expert performance and providing descriptive rationale. The framework is structured to recognize PICO-style physician queries, laying the groundwork for seamless linkage between image interpretation and evidence-based decision support. A clinician user-study confirms that the conversational interface improves interpretability and could reduce cognitive load during procedures. By uniting cutting-edge multimodal transformers with clinically curated GI data, this work advances VQA from proof-of-concept toward a practical, interactive assistant that “sees and speaks” for the endoscopist, ultimately aiming to enhance diagnostic consistency, training, and patient care.

Keywords: Visual Question Answering; gastrointestinal endoscopy; colonoscopy; transformer models; ulcerative colitis; medical AI

Dedication

This is an optional page. Use your choice of paragraph style for text on this page (**1\_Para\_FlushLeft** shown here).

To hide the heading at the top of this page, select the text and change the text colour to white.

Acknowledgements

This is an optional page. Use your choice of paragraph style for text on this page (**1\_Para** shown here).

Table of Contents

[Declaration of Committee ii](#_Toc66285129)

[Ethics Statement iii](#_Toc66285130)

[Abstract iv](#_Toc66285131)

[Dedication v](#_Toc66285132)

[Acknowledgements vi](#_Toc66285133)

[Table of Contents vii](#_Toc66285134)

[List of Tables viii](#_Toc66285135)

[List of Figures ix](#_Toc66285136)

[List of Acronyms x](#_Toc66285137)

[Glossary xi](#_Toc66285138)

[Preface/Executive Summary/Image xii](#_Toc66285139)

[Chapter 1. Introduction 1](#_Toc66285140)

**Chapter 2. Survey of VQA in Medicine with Application to Colonoscopy………13**

[References 2](#_Toc66285141)3

[Appendex 2](#_Toc66285141)5

[Appendix 3](#_Toc66285142)

List of Tables

**No table of figures entries found.**

Use References>Insert Caption to create caption labels and numbers. Right-click on the text above and select Update Field to update this list. See the Thesis Template Instructions for information on creating tables, figures, and captions.

Remember to delete this note before submission.

If there are no tables in the document, remove this page.

List of Figures

List of Acronyms

|  |  |
| --- | --- |
|  |  |
|  |  |

Glossary

|  |  |
| --- | --- |
| Thesis | An extended research paper that is part of the final exam process for a graduate degree. The document may also be classified as a project or collection of extended essays. |
| Glossary | An alphabetical list of key terms |
|  | This is an optional page and can be removed if not used. |
|  | Use one table row for each item to allow sorting using Word’s table tools. |
|  | Apply the style **1\_Para\_NoSpace** to table rows as shown here. |
|  |  |

Preface/Executive Summary/Image

This page can be used for a Preface, Executive Summary, or introductory image. This is an optional page and can be deleted if not used.

To hide the heading at the top of this page – e.g., if using an introductory image – select the text and change the text colour to white.

Do not delete the section break that follows this paragraph. If the section break is not visible, turn on non-printing characters using the Show/Hide icon (¶) on the Home ribbon.

# Introduction

**Background: Visual Question Answering in the Medical Domain**

Medical Visual Question Answering (MedVQA) is an interdisciplinary field that integrates computer vision, natural language processing, and clinical medicine 1. It aims to build AI systems capable of interpreting medical images and answering clinically relevant questions expressed in natural language.

MedVQA offers significant value for clinical decision support, with potential benefits including improved diagnostic accuracy, reduced physician workload, and broader access to expert-level insights in telemedicine and low-resource settings. Its core objective is to replicate a clinician’s ability to analyze medical images and provide meaningful answers, thereby functioning as an intelligent assistant within diagnostic workflows.

Unlike general-domain VQA, MedVQA faces distinct challenges. Medical images often contain subtle patterns or overlapping features that require domain-specific expertise to interpret accurately. Additionally, clinical questions typically demand complex reasoning, integrating visual cues with prior medical knowledge. These demands make MedVQA a rigorous benchmark for evaluating the reasoning capabilities and real-world reliability of multimodal AI systems in healthcare.

Early approaches to MedVQA were largely discriminative in nature, meaning they treated the problem as one of classification or answer selection from a fixed set of options. Many initial systems would encode the medical image (often using a convolutional neural network) and the question (using methods like word embeddings or BERT) and then predict an answer by choosing from a predefined list of possible answers. While such discriminative systems worked for constrained tasks, they struggled with open-ended or nuanced clinical questions outside the narrow answer set. For instance, if a system was only trained to answer whether a certain condition is present (“Yes” or “No”), it would be unable to handle a more detailed question like “What are the notable abnormalities in this colonoscopy image?” beyond the limited candidates provided during training. This limitation has led to a paradigm shift in the field: contemporary research is moving toward generative models that can produce free-form text answers, much like a human expert would explain an image finding. This shift has been enabled by recent breakthroughs in transformer-based Large Language Models (LLMs) and Vision–Language Models (VLMs) that are capable of open-ended reasoning. State-of-the-art MedVQA systems now exemplify this trend by combining powerful image encoders (often pretrained on medical images) with advanced language decoders that have been fine-tuned for biomedical contexts. Notable examples include Med-Flamingo and LLaVA-Med, which integrate visual processing with transformers like GPT-style decoders to generate rich, context-sensitive answers. These models represent a move toward more natural, human-like diagnostic reasoning in AI, as they can articulate explanations or insights rather than just picking an answer.

Crucially, the evolution from fixed-answer models to generative models has been paralleled by advancements in biomedical language understanding. Transformer models specialized for the medical domain have become key components in MedVQA systems. For example, BioBERT is a variant of the BERT transformer trained on biomedical text; it excels at understanding medical language and has been used for tasks like question encoding and information retrieval in healthcare applications. However, BERT-based models like BioBERT are fundamentally discriminative and not designed to generate text 2. They can classify or extract information with high accuracy but cannot natively produce free-form answers, which limits their direct use in open-ended Q&A. To address this gap, researchers developed BioGPT, a domain-specific generative transformer model pre-trained on millions of biomedical literature texts. BioGPT retains the biomedical knowledge learned from text and can generate fluent natural language – an ability that has been demonstrated with strong results on tasks such as answering questions in the PubMedQA dataset. The advent of BioGPT and similar large biomedical language models means that a MedVQA system can leverage both visual understanding and the vast textual knowledge in medicine. In practice, a MedVQA pipeline might use an image encoder to derive visual features and then use a biomedical LLM (like BioGPT) to generate an answer, ensuring that the answer is not only image-relevant but also medically coherent and informative.

Despite these technological advances, a major bottleneck for MedVQA has been the scarcity of high-quality, domain-specific training data 3. Many of the early MedVQA benchmarks were limited in size or scope. For instance, the popular VQA-RAD dataset (focused on radiology) contains only a few hundred images and a few thousand QA pairs, and mostly covers basic factual questions. Similarly, the SLAKE and PMC-VQA datasets introduced more questions but either remained small-scale or heavily skewed toward certain modalities (like X-rays and CT scans). In general, many existing QA pairs in early datasets were simplistic or even synthetically generated, which risks introducing noise and fails to test the advanced reasoning we ultimately want in clinical AI. Furthermore, traditional evaluation metrics used in VQA (such as BLEU or ROUGE scores for text similarity) often do not capture the clinical correctness or usefulness of an answer. This has underscored the need for new, richer benchmarks and evaluation methods that align with real-world clinical use – for example, scoring an answer based on whether it leads to the correct diagnosis or decision, rather than just its wording.

Within this context, the domain of gastrointestinal (GI) endoscopy has emerged as a particularly important and challenging frontier for MedVQA. GI endoscopy (which includes procedures like colonoscopy) is a routine yet critical diagnostic tool that generates large volumes of images and video. These images are visually complex – they often contain irregular shapes, fluid or debris, glare from lights, motion blur from camera movement, and variations in anatomy from patient to patient. All these factors make automated interpretation difficult, and indeed GI endoscopy images have historically received less attention in the VQA research community compared to more static imaging like radiology. Until recently, there were only a few GI-specific VQA resources. Notably, the Kvasir-VQA dataset released in 2024 was one of the first to target GI endoscopy QA, providing 6,500 endoscopic images (derived from the HyperKvasir repository) with 58,849 corresponding question-answer pairs. These questions covered categories like yes/no queries, identifying the presence of polyps, locating anatomical landmarks, naming endoscopic instruments, counting findings, and so on. The Kvasir-VQA dataset was curated with medical expert input to ensure clinical relevance, marking an important step toward practical MedVQA in endoscopy. Around the same time, the *ImageCLEFmedical 2023* challenge introduced a dataset called MedVQA-GI, focused specifically on colonoscopy images. MedVQA-GI included around 4,000 colonoscopy images with multiple QA pairs per image, featuring question types such as polyp count and location, identification of endoscopic tools, image quality issues (like specular highlights or blur), and other GI findings. This dataset was used in a competition setting to spur the development of VQA methods for GI endoscopy.

While Kvasir-VQA and MedVQA-GI broke new ground, they also highlighted the remaining gaps. The QA pairs in these datasets were often focused on relatively simple or low-level tasks – for example, “Is there a polyp in the image?” or “What color is the instrument?”. Such questions, while important, do not fully capture the depth of reasoning a gastroenterologist might use when assessing a case (e.g., evaluating disease severity or considering differential diagnoses). Furthermore, MedVQA-GI had limitations in its creation process: many answers lacked rigorous expert validation and the linguistic variety of the questions was somewhat limited (many questions followed similar templates), which could constrain the development of robust generative models. In short, by 2023 it was evident that the GI endoscopy domain needed a larger and more challenging VQA dataset to push research forward. This need has led to the development of Kvasir-VQA-x1 (2025) – a significantly expanded dataset for GI VQA. Kvasir-VQA-x1 builds upon the original Kvasir-VQA images but adds approximately 159,549 new question-answer pairs that are explicitly designed to require higher-order clinical reasoning. These new questions were generated with the assistance of LLMs and then vetted by medical experts, ensuring that they are both realistic and non-trivial. Importantly, the questions in Kvasir-VQA-x1 exhibit greater linguistic diversity and often combine multiple pieces of information, thereby testing a model’s ability to perform multi-step inference. For instance, instead of a simple one-step question like “How many polyps are present?”, a complex question might be “Describe all notable findings and their locations, and suggest a potential diagnosis.” Such questions demand that a model understand various elements in the image and synthesize an answer – a much more rigorous evaluation of AI capabilities.

Additionally, Kvasir-VQA-x1 introduces a series of visual perturbations to the images (such as occlusions, contrast changes, and blurring) to simulate real-world variability and to test the robustness of VQA models. Each question-answer pair in the dataset is also annotated with a complexity level that quantifies the difficulty of the question (in terms of required reasoning and image analysis). This allows researchers to evaluate how models perform on easy vs. hard questions, and whether they can handle incrementally more complex scenarios. By providing both standard and perturbed versions of images (in a dual-track evaluation), Kvasir-VQA-x1 enables comprehensive benchmarking of not only accuracy but also resilience of MedVQA systems. In summary, the field of medical VQA has rapidly progressed from simple QA pairs and classification-based models to a new generation of systems that leverage transformer-based generative AI, are trained on large-scale curated datasets, and aspire to truly assist clinicians by answering meaningful, complex questions about medical images.

**Motivation for VQA in Gastrointestinal Endoscopy (Colonoscopy)**

Gastrointestinal endoscopy, and colonoscopy in particular, is a domain of medicine that stands to benefit enormously from advances in visual question answering. Colonoscopy is widely used for both the diagnosis and management of diseases in the colon (large intestine), including inflammatory conditions like ulcerative colitis and Crohn’s disease, as well as for the detection and removal of precancerous polyps. As a procedure, colonoscopy produces a rich stream of visual data, typically hours of video or numerous high-resolution images per session, which must be interpreted by specialists in real-time and afterwards. In current practice, AI tools have already begun to assist endoscopists by performing tasks like polyp detection, localization, and sizing, but these tools usually present their outputs as simple overlays, binary signals, or segmentations (for example, drawing a bounding box around a detected polyp or outputting a severity score number) 3. While useful, such outputs are not necessarily user-friendly or informative to a clinician beyond the immediate detail. A numeric score or a highlighted region does not explain why the model made a certain assessment or what other context might be relevant. This is where VQA offers a more natural and interpretable interface: a VQA system can engage in dialogue, answering questions like “Do you see any polyps?,” “How large is the polyp in the sigmoid colon?,” or “Is there active bleeding or ulceration present?” in human language. By doing so, the AI’s findings are communicated in the same terms a doctor would use in their own notes or conversations, potentially making the AI’s assistance more transparent and easier to trust. Research in medical AI suggests that such natural language interaction can improve the interpretability of the system’s output and thereby increase clinician acceptance of AI tools. In essence, VQA could transform an endoscopy AI from a silent image analyzer into an interactive “second pair of eyes” that a physician can consult during procedures or afterward.

The motivation to apply VQA specifically to colonoscopy is further strengthened by the clinical demands and challenges surrounding diseases like ulcerative colitis (UC). UC is a chronic inflammatory bowel disease that affects the colon, and its management heavily relies on endoscopic evaluation of disease activity. Gastroenterologists use scoring systems such as the Mayo Endoscopic Score (MES) or the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) to grade how severe the inflammation is during a colonoscopy exam. These scores range, for example, from 0 (normal mucosa, no inflammation) up to 3 (severe inflammation with ulcers and spontaneous bleeding in the colon). Grading the severity is crucial: it guides treatment decisions, such as intensifying therapy for high scores or maintaining course for low scores, and it serves as an indicator of prognosis. However, visual grading is also subjective and requires significant expertise. Automating this process with AI has been a goal in recent years, and indeed deep learning models have shown excellent performance in classifying endoscopic images by UC severity, often matching the accuracy of experienced human endoscopists. For instance, convolutional neural network models have been trained to distinguish between Mayo 0, 1, 2, and 3 with impressive accuracy, and some studies report that AI predictions of “mucosal healing” or active disease correlate well with what gastroenterologists would conclude. This is not only academically interesting but practically important: if AI can provide real-time grading of disease severity during colonoscopy, it could help ensure that treatment decisions are made on the most objective basis and potentially reduce inter-observer variability among doctors.

Integrating such capabilities into a VQA system is a logical next step. Instead of just outputting a severity grade or a probability, a VQA-based approach would allow the system to answer specific questions about disease activity in descriptive terms. A physician could ask, for example, “How severe is the inflammation in this region of the colon?” and the system might answer: “It appears to be severe ulcerative colitis with deep ulcerations and bleeding, consistent with a Mayo score of 3.” This kind of answer not only provides the score but also justifies it by describing the visual findings (ulcerations, bleeding) that lead to that conclusion. Such justifications can greatly increase a clinician’s confidence in the AI’s assessment by showing that it “noticed” the same relevant features a human would. Moreover, the ability to query the system in natural language means the doctor can probe the AI further: “Do you see any areas of normal mucosa or is it all inflamed?” or “Are there any complications like strictures or pseudopolyps visible?”. In this way, a VQA system in colonoscopy becomes a conversational assistant that can discuss the image findings, much like a trainee or a colleague would during a procedure.

The need for such intelligent assistance in colonoscopy is underscored by broader healthcare trends. The incidence of inflammatory bowel disease (which includes UC) has been rising, and managing these chronic patients requires frequent endoscopic assessments. At the same time, healthcare systems worldwide face a shortage of specialized clinicians. In the United States, for example, projections indicate a shortfall of roughly 1,600 gastroenterologists by the year 2025. This gap between demand and supply means that many patients might not receive timely colonoscopies or detailed attention to their endoscopic findings, potentially delaying important management decisions. AI-driven tools, including MedVQA systems, offer a way to ameliorate the strain on specialists by automating parts of the diagnostic process. If a VQA system can reliably answer routine questions about a colonoscopy (identifying normal vs. abnormal findings, grading severity, counting polyps, etc.), the physician can devote more focus to complex decision-making and patient communication. In settings with less experienced clinicians, an AI assistant could serve as a real-time tutor or double-check, thus elevating the standard of care and reducing diagnostic errors. Studies have suggested that AI assistance can reduce human error rates and improve consistency in image-based diagnoses, which is particularly pertinent in interpreting endoscopic imagery that may be prone to variability in human judgment.

Another important motivation for incorporating VQA in the colonoscopy workflow is its potential role in clinical decision support beyond image interpretation. Often, a finding on a colonoscopy prompts further clinical questions that go beyond the image itself and into the realm of medical knowledge. For example, if a severe inflammation is observed (Mayo 3), the physician might wonder about optimal treatment adjustments: “In a patient with severe ulcerative colitis unresponsive to current therapy, what are the next treatment options?” While answering such a question fully requires knowledge of medical guidelines and evidence (not just the image), a sophisticated VQA system could be connected to a knowledge base or leverage a medical LLM to provide useful information. Here, drawing on the principles of evidence-based medicine is key. Clinicians often formulate questions about patient care in terms of PICO elements – Patient/Problem, Intervention, Comparison, Outcome – to systematically search for evidence. For instance, a PICO-style question might be, “In a patient with moderate ulcerative colitis (P), how effective is adding biologic therapy (I) compared to continuing standard mesalamine (C) for inducing remission (O)?” A truly advanced MedVQA system might interpret a question about an image in a similar structured way, recognizing the patient/problem in the image (e.g., “moderate ulcerative colitis” from the colonoscopy) and connecting that with intervention/outcome queries. In fact, recent research by Mohammed and Fiaidhi (2024) has shown the feasibility of combining domain-specific language models with PICO-aware frameworks to answer clinicians’ queries: they used a bootstrapped ensemble of BioBERT and BioGPT to better understand physician questions structured in PICO format and retrieve answers from medical literature. This demonstrates how integrating an image-based system with text-based evidence retrieval could allow a physician to ask comprehensive questions like, “Given the endoscopic findings of severe colitis, what do guidelines suggest as the next step in management?” and receive an answer that merges the visual assessment with current medical knowledge (potentially citing clinical trial data or guidelines). While the primary focus of this thesis is on VQA within the scope of interpreting medical images, being mindful of such extensions is important. It highlights that VQA in medicine is part of a continuum from image understanding to decision support. By designing our system with the capability to handle free-form questions, we lay the groundwork for future integration with evidence-based query answering protocols such as PICO. In summary, the motivation for applying VQA to colonoscopy images is not only to improve image interpretation (making it more interactive and accurate), but also to situate that interpretation within the larger context of patient care, thereby directly supporting clinical decisions.

**Research Objectives**

Based on the background and motivations outlined, this thesis sets out the following research objectives:

**Objective 1: Develop a MedVQA system for gastrointestinal endoscopy (colonoscopy)** – Design and implement a visual question answering framework capable of interpreting colonoscopy images and generating accurate, clinically relevant answers to questions. This includes selecting or creating appropriate vision and language models (e.g., convolutional or transformer-based image encoders coupled with biomedical language models) and integrating them into an end-to-end VQA pipeline for the medical domain.

**Objective 2: Leverage both classic and generative AI approaches** – Explore the combination of discriminative (classification-style) techniques and generative transformer-based techniques for medical VQA. In particular, utilize transformer models (such as BioBERT for question understanding and BioGPT or related LLMs for answer generation) and assess how domain-specific pre-training can improve performance in answering medical questions. The system should be able to handle both closed-form questions (e.g., yes/no, multiple-choice) and open-ended questions requiring free-text answers.

**Objective 3: Utilize and contribute to state-of-the-art MedVQA datasets** – Make use of existing datasets like Kvasir-VQA and MedVQA-GI to train and evaluate the system, and incorporate the newly released Kvasir-VQA-x1 dataset to push the model’s reasoning capabilities. If necessary, extend these datasets (for example, by additional annotation or creating new question types) to cover specific aspects of colonoscopy that are important (such as detailed descriptions of ulcerative colitis severity). Ensuring the dataset used for training/evaluation has a wide coverage of clinically relevant question types is a key part of this objective.

**Objective 4: Integrate clinical decision support elements** – Ensure that the VQA system’s design accounts for practical clinical needs, particularly ulcerative colitis severity grading and related decision support. This means the system should be able to answer questions that a gastroenterologist might ask during a colonoscopy (e.g., “Is there a severe inflammation? What is the likely Mayo score?”) and the answers should align with standard clinical criteria. Additionally, prepare the system to handle physician queries in a structured way; for instance, by recognizing when a question implies a need for external knowledge (following the PICO framework) and structuring the answer or further action accordingly. This objective will likely involve linking the vision-based answers with text-based information (such as guidelines or literature) in a seamless manner.

**Objective 5: Evaluate the system’s performance and contributions** – Rigorously evaluate the developed VQA system on multiple fronts: accuracy of the answers (compared to ground truth annotations or expert answers), robustness to image perturbations and variability, and usefulness in a clinical context. The evaluation should include standard quantitative metrics for VQA, but also qualitative analysis of example Q&A outputs, especially for complex questions. Where possible, expert assessment (e.g., having medical professionals judge the correctness and helpfulness of the AI’s answers) will be incorporated. This objective ensures that the research does not stop at algorithm development but also measures the real-world applicability of the proposed solution.

Through these objectives, the thesis aims to bridge the gap between advanced AI models and practical clinical utility in the realm of endoscopic image analysis.

**Research Questions**

To guide the investigation, the following central research questions are posed:

How can generative visual question answering techniques be effectively applied to medical images from gastrointestinal endoscopy? This question examines the adaptation of multimodal AI models (vision + language) to the context of colonoscopy. It probes what model architectures and training strategies are needed for the AI to interpret endoscopic images and answer domain-specific questions accurately.

What are the benefits of a VQA-based interface compared to traditional computer-aided diagnosis outputs in colonoscopy? Here we ask whether allowing a system to answer questions in natural language (versus outputting, say, a binary finding or a numeric score) improves the interpretability, completeness, or clinical usefulness of the AI’s output. For example, does a VQA system provide better explanatory support for ulcerative colitis severity grading than a simple classification model would?

Can the integration of domain-specific language models (like BioGPT/BioBERT) and knowledge frameworks (like PICO) enhance the quality of answers in a medical VQA system? This research question aims to uncover the impact of incorporating biomedical knowledge and structured clinical reasoning into the VQA process. It asks whether a system that is aware of medical context (through pre-trained biomedical transformers or PICO elements) can answer questions more accurately and with appropriate clinical justification, especially for open-ended questions that go beyond visible image features.

How robust and generalizable is the MedVQA system when faced with real-world variability in endoscopic images and questions? We seek to evaluate the system’s performance not only on standard, high-quality images and straightforward questions, but also under more challenging conditions. This includes assessing robustness to common endoscopy image artifacts (blur, poor lighting, obstructions) and to linguistically varied or complex questions that may combine multiple sub-questions. Essentially, this question addresses the system’s reliability in practical deployment scenarios and its ability to handle the breadth of cases and queries a clinician might actually encounter.

In what ways can a MedVQA system support clinical decision-making in gastroenterology, and what are the limitations? This question looks at the end-use of the technology. It asks how the developed system can be utilized in a clinical workflow (for example, as a diagnostic assistant during procedures or as an educational tool for trainees). At the same time, it prompts an analysis of the system’s limitations – whether in types of questions it cannot handle, scenarios where it might be less accurate, or the extent to which clinicians can trust and act on its answers. Identifying these limits is crucial for defining the scope of safe and effective use of the system.

By addressing these research questions, the thesis will cover the development of the VQA system, its technical performance, and its practical implications, thereby providing a comprehensive exploration of visual question answering in the context of medical colonoscopy images.

In summary, this thesis expands the horizons of Visual Question Answering in medicine by applying and enhancing it for gastrointestinal endoscopy, particularly colonoscopy. Through the development of a new VQA system, utilization of cutting-edge AI models, and focus on a pressing clinical application (ulcerative colitis management), it provides both academic and practical contributions. The work not only advances methodological understanding in multimodal medical AI but also delivers a step toward AI systems that can converse with clinicians about medical images, thereby supporting decision-making in a transparent and evidence-aligned manner. The results and contributions herein set the stage for more interactive and intelligent clinical support tools in the near future, aligning with the ultimate goal of improved patient care through technology. Throughout this chapter, we have established the background, motivations, and aims of the research. In the following chapters, we will delve into the technical foundations, methodology, experimental results, and discussions that build upon this introduction, focusing on the implementation details and outcomes of the proposed MedVQA system for colonoscopy. The overarching vision is clear: enabling machines to see and speak about what they see in a way that adds value to medical professionals and their patients, and this thesis takes significant steps toward realizing that vision in the context of gastrointestinal endoscopy.

# Survey of Visual Question Answering in Medicine with Application to Colonoscopy

Visual Question Answering (VQA) in medicine is an emerging field that integrates computer vision and natural language processing to enable AI systems to interpret medical images and answer clinically relevant questions 5. Unlike general-domain VQA, medical applications require expert-level image understanding and the ability to reason over complex clinical context. Recent advances in deep learning and multimodal models have expanded the potential of VQA across various medical imaging domains, including radiology and endoscopy. Among these, colonoscopy stands out as a key area where VQA can support clinicians in polyp detection, disease assessment, and decision-making. This survey focuses on VQA techniques and datasets specifically relevant to colonoscopy.

**VQA in Medical Imaging: Early Efforts and Challenges**

Medical VQA research began in the late 2010s with the introduction of the first curated datasets combining medical images and question-answer (QA) pairs. Notably, VQA-RAD 6 was the first manually constructed radiology VQA dataset, featuring clinician-generated questions about radiographs. Around the same time, the ImageCLEF 2018 7 challenge introduced a VQA task based on radiology images, helping spark broader interest in the field. These early datasets were limited in scale, VQA-RAD included only 315 images and ~3,500 QA pairs, and focused primarily on modalities like X-rays and CT scans. Despite their size, they demonstrated the feasibility of VQA in clinical domains and surfaced critical challenges: medical images often contain subtle visual cues, and many questions require domain-specific reasoning beyond the image itself.

Early VQA models treated the task as a classification or multiple-choice problem, performing reasonably well on constrained question types like yes/no or modality identification. However, they struggled with open-ended, free-form questions common in clinical practice. Additionally, many early datasets contained biased or overly simplistic QA pairs, allowing models to rely on shallow patterns rather than genuine understanding. These limitations prompted the development of more diverse datasets and more advanced modeling approaches capable of handling the complexity and high-stakes nature of medical VQA.

Table 1: Key medical VQA datasets, showing the progression from general radiology datasets to domain-specific gastrointestinal (GI) endoscopy datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Year/Modality** | **Images/QA Pairs** | **GI Relevance/Notes** |
| **VQA- RAD** | 2018 / Radiology | 315 / 3,500 | First medical VQA dataset; not GI-specific |
| **ImageCLEF VQA-Med** | 2018-2019 / Radiology | N/A | Community challenge; general radiology |
| **SLAKE** | 2021 / CT, MRI, X-ray | 642 / 14,000 | Synthetic QA; no GI focus |
| **PathVQA** | 2020 / Pathology | N/A | Pathology slide QA; microscopic images |
| **PMC-VQA** | 2023 / Mixed modalities | ~149,000 / 227,000 | Text-mined QA from publications; limited GI cases |
| **MedVQA-G** | 2023 / Colonoscopy | ~4,000 / Multiple per image | First GI-focused VQA dataset; polyp count, tool detection |
| **Kvasir-VQA** | 2024 / GI Endoscopy | 6,500 / ~58,849 | Full GI tract (esophagus to colon); disease ID, count, location, etc. |
| **Kvasir-VQA-x1** | 2025 / GI Endoscopy | N/A | Reasoning-oriented QA; expanded version of Kvasir-VQA |

**Advances in Generative Vision-Language Models for Medical VQA**

Recent breakthroughs in large language models (LLMs) and vision-language models (VLMs) 8 have shifted medical VQA from answer selection to free-text answer generation. Unlike earlier systems limited to fixed choices, generative models produce context-aware, descriptive responses that resemble expert clinical explanations.

State-of-the-art architectures typically combine pre-trained image encoders (e.g., CNNs, ViTs) with instruction-tuned LLM decoders adapted for medical dialogue. Notable examples include Med-Flamingo, LLaVA-Med, and MedGEM, which pair CLIP-like vision encoders with domain-specific language models. These systems are further fine-tuned on image-text datasets rich in clinical terminology and visual nuance.

To adapt general-purpose models to medical VQA tasks, researchers employ parameter-efficient techniques like prefix tuning, LoRA adapters, and visual prompt tuning. These methods allow fine-tuning on limited data while avoiding overfitting. Some models also integrate biomedical knowledge graphs (e.g., UMLS) to improve reasoning over medical concepts.

Common design strategies across top-performing generative MedVQA systems include:

1. Aligning visual and textual features through learned projections.
2. Pre-training on medical image-caption pairs.
3. Instruction tuning with curated clinical QA data.
4. Injecting domain knowledge via adapters or additional inputs.

These advances enable models to generate paragraph-level answers, express uncertainty, and provide reasoning steps, capabilities far beyond earlier classification-based systems.

However, challenges persist. Generative models may hallucinate plausible but incorrect responses when faced with unfamiliar queries. This raises concerns around factual reliability and clinical safety. To address this, researchers are exploring evaluation methods beyond string-matching, such as semantic similarity scoring and clinician-based review.

Despite these challenges, generative VQA represents the future of medical image interpretation, aligning closely with how clinicians naturally communicate - through free-form, nuanced explanations grounded in visual evidence.

**Visual Question Answering for GI Endoscopy and Colonoscopy**

Colonoscopy and related GI endoscopic procedures have become key focus areas for medical VQA due to the need for AI assistance in lesion detection, anatomical identification, and procedural interpretation. While earlier AI systems targeted polyp detection or segmentation, VQA enables a broader scope, answering targeted clinical questions such as “How many polyps are present?”, “Where is the abnormality?”, or “What tool is visible in the frame?”

A major milestone was the ImageCLEFmed 2023 MedVQA-GI challenge, the first public competition on VQA for GI endoscopy images 10. It provided a dataset of ~5,000 colonoscopy and gastroscopy images, each paired with multiple clinically guided QA pairs. The challenge included:

* VQA (answering questions about an image),
* Visual Question Generation (VQG) (generating a question given an image and answer),
* Visual Location QA (VLQA) (segmenting the region that answers the question).

Common tasks included polyp detection, counting, anatomical labeling, instrument identification (e.g., snares, forceps), and quality assessment. Despite some answers lacking expert validation or being subjective (e.g., image quality grading), the competition drove innovation, with top models combining vision encoders and transformer-based question understanding, sometimes enhanced with medical knowledge or ensembling strategies.

Building on this, Kvasir-VQA (2024) was introduced as a comprehensive dataset with 6,500 GI endoscopy images and ~58,849 QA pairs. It covers the full GI tract (esophagus to colon) and supports multiple question types: yes/no, location-based, counting, multiple-choice, and color-based queries. The dataset also supports tasks like image captioning, report generation, and text-to-image generation, enabled by rich annotations sourced from HyperKvasir and Kvasir-Instrument.

To advance beyond basic visual pattern recognition, Kvasir-VQA-x1 (2025) was introduced with an emphasis on reasoning-intensive VQA. This extension includes more complex questions involving multi-image reasoning or clinical judgment, e.g., “Based on these findings, what is the likely diagnosis and recommended follow-up?” The dataset aims to benchmark higher-order reasoning and encourages clinician-guided evaluation metrics beyond traditional accuracy or BLEU scores.

Collectively, these datasets and challenges underscore the unique demands of GI endoscopy VQA - ranging from visual artifacts and fine-grained lesion detection to clinical decision support. As models evolve, the goal is to develop AI assistants capable of interpreting endoscopic scenes and answering diagnostic questions in real-time, improving clinical workflow and decision-making.

**Ulcerative Colitis Severity Assessment with Generative AI**

Ulcerative colitis (UC), an inflammatory bowel disease, is assessed via colonoscopy using standardized scoring systems such as the Mayo Endoscopic Subscore (MES) and the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). Automating this scoring process supports objective monitoring of disease progression and treatment response.

Traditional AI models have treated severity assessment as a classification task, where a colonoscopy image is mapped to a discrete score (e.g., MES 0–3). For instance, one study achieved an F1-score of 0.92 when distinguishing between MES 0 and 1, surpassing novice endoscopists in consistency. Other models trained on colonoscopy videos have shown ~90% accuracy in predicting UCEIS scores aligned with expert ratings 10. While these systems perform well numerically, they lack interpretability. Generative AI techniques aim to address this by producing both the severity score and a descriptive explanation, such as: “Moderate colitis (Mayo 2) with marked erythema and partial mucosal erosion.” This mirrors how gastroenterologists document findings in clinical reports.

One approach integrates Siamese neural networks to compare new cases against prototypical examples of each severity grade, supporting few-shot learning in small datasets. By incorporating “thick data” (clinical context, metadata, expert-driven features), these models improve classification accuracy and generate more informative outputs 11. Additional methods involve caption generation guided by explainable AI (xAI), which highlights visual regions (e.g., ulceration, bleeding) that contributed to the assessment. This dual output - classification and rationale, enhances transparency and may support use in automated reporting and training scenarios. General-purpose large language models (LLMs) have also been tested for UC scoring. One review found that LLM-based severity classifications matched clinician assessments approximately 80% of the time. However, concerns about factual consistency and reliability persist, especially when deployed without domain adaptation.

Overall, UC severity assessment is transitioning from pure classification to generative, explainable systems that combine VQA, medical language generation, and visual reasoning. These systems hold promise for improving clinical documentation, decision support, and consistency in disease evaluation.

**Generative AI for Physician Queries Using the PICO Framework**

In clinical practice, physicians often pose complex management questions, e.g., “In ulcerative colitis patients with moderate severity, is Drug A more effective than Drug B for inducing remission?” These queries are typically addressed through evidence-based medicine and literature review. Structuring such questions using the PICO format (Patient/Problem, Intervention, Comparison, Outcome) enables focused information retrieval and reasoning.

Recent work in generative AI has explored using PICO-aware systems to automate evidence synthesis. For example, PICO GenAI 11 is a system that uses PICO-formulated queries to retrieve relevant clinical literature and generate coherent summaries using large language models (LLMs). The output resembles a clinical report, citing data from trials, e.g., remission rates or adverse events - based on the specified population, interventions, and outcomes.

A related approach by 12 applied retrieval-augmented generation (RAG) with GPT-4 to answer questions about clinical trial eligibility. The system retrieved relevant text segments from trials or guidelines and used the LLM to compose grounded, evidence-based answers.

Such methods are directly applicable to ulcerative colitis. For instance, a physician might ask, “What are the treatment options for steroid-refractory moderate UC?” The system identifies PICO elements (e.g., P = moderate UC, I = second-line therapies, O = remission) and retrieves data from sources like ACG or ECCO guidelines to synthesize a response. This allows for precise differentiation between induction vs. maintenance therapy, adult vs. pediatric populations, and other clinically relevant variables.

To ensure reliability, these systems are designed to cite sources explicitly and minimize hallucinations through fact-grounding. A pilot study on PICO GenAI showed that such systems could generate helpful summaries from trials and case reports. Likewise, LLMs like ChatGPT have shown ~80% alignment with clinical assessments in tasks such as disease severity classification and clinical Q&A, though consistency and factual correctness remain active concerns.

In the context of colitis, these tools could support queries about management strategies, prognosis, or therapy sequencing. For example, “What evidence supports JAK inhibitors in patients unresponsive to TNF blockers?” The system could retrieve and summarize relevant studies, supporting the physician’s decision-making.

In summary, generative AI applied to structured physician queries represents an extension of VQA beyond images into text-based clinical decision support. These systems combine natural language understanding, medical information retrieval, and evidence-grounded generation to respond to PICO-structured questions. While still under development, they offer potential for real-time, literature-backed assistance during clinical care, particularly in areas like gastroenterology where treatment decisions are nuanced and guideline-driven.

**Conclusion**

Visual Question Answering (VQA) in the medical domain, particularly for colonoscopy and gastrointestinal (GI) imaging - has progressed from foundational work in radiology to advanced applications in endoscopy. This evolution is driven by the availability of specialized datasets (e.g., Kvasir-VQA) and the integration of multimodal generative AI, combining vision encoders with large language models (LLMs) to enable more natural, context-aware answers.

Modern VQA systems can now perform tasks such as polyp counting and localization, tool identification, and generation of natural-language summaries. In parallel, clinical applications like ulcerative colitis severity assessment have benefited from both classification models (for accurate scoring) and generative models (for interpretability and report-like outputs). These developments aim to make AI not only accurate but also clinically interpretable and useful in decision-making.

Bibliography

1. Sushant Gautam, Michael A. Riegler, and Pål Halvorsen, “Kvasir‑VQA‑x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy,” arxiv.org (preprint, June 11, 2025), <https://ar5iv.labs.arxiv.org/html/2506.09958v1,>
2. Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. “BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining.” arXiv.org, October 19, 2022. <https://arxiv.org/abs/2210.10341>.
3. “Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract,” CEUR Workshop Proceedings, no. 3497, Paper 107 (n.d.), accessed July 18, 2025, <https://ceur-ws.org/Vol-3497/paper-107.pdf>.
4. “Diagnosis and Severity Assessment of Ulcerative Colitis Using Self Supervised Learning.” Accessed July 18, 2025. <https://arxiv.org/html/2412.07806v1>.
5. Gautam, Riegler, and Halvorsen, “Kvasir‑VQA‑x1,” [1]
6. Lau, Jessie Y. C., James R. Glass, and H. Chen. "A Dataset and Exploration of Models for Understanding Radiology Reports via QA." *Proceedings of NAACL-HLT*, 2018. <https://aclanthology.org/N18-2094/>
7. Lau, J., Gayen, S., Ben Abacha, A. et al. A dataset of clinically generated visual questions and answers about radiology images. Sci Data 5, 180251 (2018). <https://doi.org/10.1038/sdata.2018.251>
8. Dong, Wenjie, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. 2025. “Generative Models in Medical Visual Question Answering: A Survey.” Applied Sciences 15, no. 6: 2983.<https://doi.org/10.3390/app15062983.>
9. **ImageCLEFmed MEDVQA‑GI. 2023.** ImageCLEF / LifeCLEF – Multimedia Retrieval in CLEF. Accessed July 18, 2025. <https://www.imageclef.org/2023/medical/vqa>.
10. Murino, Alberto, and Alessandro Rimondi. 2023. “Automated Artificial Intelligence Scoring Systems for the Endoscopic Assessment of Ulcerative Colitis: How Far Are We from Clinical Application?” Gastrointestinal Endoscopy 97 (2): 347–49. <https://doi.org/10.1016/j.gie.2022.10.010>.
11. Mohammed, Sabah, and Jinan Fiaidhi. 2024. “Generative AI for Evidence‑Based Medicine: A PICO GenAI for Synthesizing Clinical Case Reports.” In Proceedings of the IEEE International Conference on Communications (ICC 2024), June 9–13, 2024, Denver, CO, USA. <https://doi.org/10.1109/ICC51166.2024.10622271>.
12. Xu, Zihan, Haotian Ma, Gongbo Zhang, Yihao Ding, Chunhua Weng, and Yifan Peng. 2025. “Natural Language Processing in Support of Evidence‑based Medicine: A Scoping Review.” *arXiv* (Preprint), May 28, 2025. <https://arxiv.org/abs/2505.22280>.

Appendix A.

Bibliography

“Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract,” n.d. Accessed July 16, 2025.

Bibliography

ar5iv. “Kvasir-VQA-X1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy,” n.d. https://ar5iv.labs.arxiv.org/html/2506.09958v1.