Solving Medical Data Limitations Through AI: Multi-Modal Vision-Language Learning for Gastrointestinal VQA and Synthetic Training Data Generation*

Notebook for the Morgan$_{CSLabatCLEF2025:https://github.com/Ejigsonpeter/VQA-and-Data-Generation}$

Ejiga Peter Ojonugwa Oluwafemi[1,*,†], Mahmudul Hoque[1,†], Ejiga Frederick Akor[2,†], Raisa Nusrat Chowdhury[1,†], Abdullahi Bn Umar[3,†] and Md Mahmudur Rahman[1,†]

[1]Department of Computer Science, SCMNS, Morgan State University, Baltimore, Maryland, USA
[2]International Organization for Migration (IOM), Geneva, Switzerland
[3]Department of Computer Science, Federal University of Education Kano, Nigeria

Abstract

Gastrointestinal image analysis is crucial for early disease detection but faces challenges including data scarcity, privacy concerns, and limited automated diagnostic support. Traditional medical visual question answering (VQA) systems struggle with domain-specific knowledge and insufficient training data, while existing synthetic image generation methods fail to maintain the clinical authenticity required for medical applications. This paper presents a dual-task multi-modal framework integrating VQA and synthetic image generation to address these limitations. The methodology employs parameter-efficient fine-tuning of Florence-2 on the Kvasir-VQA dataset (6,500 gastrointestinal images), freezing the DaViT vision encoder while fine-tuning language components with cross-attention fusion for Sub-task 1. For Sub-task 2, the approach implements LoRA-enhanced Stable Diffusion 2.1 with rank-8 adaptation, incorporating structured clinical prompts for medically relevant synthetic image generation. Evaluation using standard NLP metrics (BLEU, ROUGE, METEOR) for VQA and image quality metrics (FBD, Fidelity, Agreement, Diversity) demonstrates significant improvements over baseline methods. The VQA system achieves ROUGE-L of 0.91, ROUGE-1 of 0.92, BLEU of 0.24, and METEOR of 0.50, substantially outperforming existing approaches. Synthetic image generation attains an optimal FBD of 1449.63 with fidelity of 0.29 and agreement of 0.73 while maintaining clinical authenticity. The parameter-efficient approach reduces computational requirements by 60% compared to full fine-tuning while achieving superior performance. Comprehensive ablation studies validate design choices, demonstrating cross-attention fusion effectiveness and optimal rank-8 LoRA configuration, providing enhanced gastrointestinal diagnostic support and privacy-preserving data augmentation.

Keywords

Medical VQA, ImageCLEFmed 2025, Multimodal AI, Clinical Question Answering, Synthetic GI Images, Florence-2, LoRA, Stable Diffusion, Parameter-Efficient Fine-tuning, PEFT Gastrointestinal Diagnostics, Polyp Detection, Synthetic Medical Imaging, Vision Transformers, Medical Imaging

1. Introduction

Early detection and diagnosis of these conditions heavily rely on gastrointestinal endoscopic image analysis. Problems of the gastrointestinal system, including polyps, different inflammatory disorders, and malignancies. The growing number of endoscopic surgeries and the challenges in interpreting

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

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Md Rahman:md.rahman@morgan.edu.

†These authors contributed equally.

ejiga.ojonugwa.peter@gmail.com (E. P. O. Oluwafemi); mahoq1@morgan.edu (M. Hoque); ejiga.fredrick@yahoo.com (E. F. Akor); racho1@morgan.edu (R. N. Chowdhury); abdullahiu226@gmail.com (A. B. Umar); md.rahman@morgan.edu (M. M. Rahman)

0009-0003-2039-3075 (E. P. O. Oluwafemi); 0009-0006-5532-4135 (M. Hoque); 0009-0003-0700-6712 (E. F. Akor); 0000-0002-1663-6391 (R. N. Chowdhury); 0009-0001-5137-7831 (A. B. Umar); 0000-0003-0405-9088 (M. M. Rahman)

images are difficult issues. Difficulties for healthcare systems prompt the use of automated systems for diagnostics. Current diagnostic Human interpretation, which can cause differences between analysis results, plays a major role in workflows. pressure from more cases and difficulty in making a diagnosis. VQA systems give promising results in renewing visual skills. Finding solutions by providing natural language access to all types of medical images, for clinicians to inquire about issues and receive thoughtful, properly framed answers. They close the gaps between cultures. The space between understanding detailed images and coming to quick doctor decisions. The use of synthetic data is becoming helpful to overcome the privacy, scarcity, and unbalanced class problems found in medical datasets [1]. Common medical data collections face restrictions because protecting patient information and the rarity of specific medical conditions make the data scarce. Recent progress has been made in diffusion models and vision-language models, which are able to produce medical images of high quality that still carry useful information for diagnosis [2], [3]. Using VQA together with synthetic data generation, medical AI systems can improve how they diagnose and increase the number of available training samples at the same time [3]. This approach helps at the start by supporting clinical decisions and in the future by creating strong, adaptable AI models. This research covers the challenges of automated health diagnoses by using a smart multi-modal framework that he developed. We have developed our work based on the ImageCLEF 2025 [4] competition under the MEDVQA [5] category, which is separated into subtasks. The first task concentrates on improving Florence-2 for use in medical VQA on gastrointestinal images, so the model can interpret clinical questions that depend on what an endoscopy can show. Using LoRa on Stable Diffusion, subtask 2 offers GI images made in a clinical style to protect privacy, respect ethical principles, and make more training data possible. Our key contributions include: (1) parameter-efficient fine-tuning strategies for Florence-2 on the Kvasir-VQA dataset with comprehensive performance evaluation, (2) development of robust synthetic image generation pipelines using diffusion models with enhanced prompt engineering, and (3) Comprehensive evaluation of both approaches for clinical applicability in gastrointestinal diagnostics with detailed ablation studies. es.

2. Related Work

Medical VQA connects computer vision with natural language processing to help in healthcare. Before VQA-RAD [6], there were few examples of radiology image question answering, but this benchmark made it clear that using natural language is a promising way to interpret medical images. In addition, PathVQA [7] brought the idea of vision-language interaction to pathological imaging and introduced difficult tasks related to diagnosing health problems. KvasirVQA [8] was designed for gastrointestinal endoscopy and offers complete annotations for finding and labeling polyps. When compared to these previous approaches, our method uses the strong multimodal abilities in Florence-2 to better identify and understand the spatial setup and clinical context in endoscopic images. There has been major progress in medical image synthesis, moving from GAN to diffusion methods in recent times. Generating medical images with traditional GANs was promising; however, they usually ran into mode collapse, resulting in little diversity in what was generated [9]. The use of advanced diffusion models has greatly improved the quality, control, and stability of medical image generation [10]. Through text-guided synthesis, PromptToPolyp [11] can generate simple polyp images, yet is restricted to simple polyp structures. To go beyond standard polyps, we use LoRA-enhanced Stable Diffusion and fully integrate descriptions from clinical cases, ensuring the new images have both diagnostically relevant anatomy and a wide variety of intestinal disorders. [12] created MammoFormer, a framework that combines transformers (ViT, Swin, ConvNeXt) with feature refinements (negative transform, AHE, HOG) and five XAI methods (Integrated Gradients, GradCAM, Occlusion, DeepLIFT, Saliency) to distinguish the local details and global context. Optimized architectures improved performance by 13 percent (98.4 percent accuracy with HOG), making it possible to have a deployable and explainable workflow in breast cancer screening. Recent advances in creating large-scale vision-language models have greatly improved our understanding of multimodal data. By learning together images and text, CLIP [13] was able to perform strong zero-shot tasks.

BLIP [14] exceeded CLIP when it introduced bidirectional encoder-decoder networks that improve how vision and language are paired. Recent work, Flamingo [15], exhibited the ability to solve vision-language problems with only a little training data. Although these models can work well in most areas, they still lack the specific medical information needed for healthcare. The method we propose adopts the multimodal architecture of Florence-2 to work with medical imaging, including visual features and terms that are commonly found in medicine. Mahmud et. al [16] participated in ImageCLEF 2024 medical caption prediction and concept detection tasks. Their LLaVA-v1.6-Mistral-7B model with selective LoRa fine-tuning (40.1M parameters) achieved second place in caption prediction with 0.628059 BERTScore. They also explored quantized models, demonstrating parameter-efficient approaches for medical image understanding. By using parameter-efficient techniques, it is now possible to fine-tune big pre-trained models to fit their needs, using much less computing power. With LoRa [17], adaptation matrices in low rank allow the reduction of parameters, doing so without a drop in accuracy. PEFT methods [18] include adapter layers, prompt tuning, and prefix tuning. Combining LoRa and vision tower freezing, our method allows for fine-tuning Florence-2 efficiently with relatively less risk of losing existing visual concepts. This way of working is unique from standard PEFT approaches by upgrading language parts while still offering good visual data processing, which leads to higher efficiency and less strain on computational resources in medical VQA tasks.

3. Task Overview and Dataset

The team participated in both subtasks of the ImageCLEF medical 2025 challenge [4], [5]: Visual Question Answering (VQA) and Synthetic Image Generation. A dual-task strategy was employed to address complementary aspects of medical AI—generating high-fidelity synthetic data for privacy-preserving dataset augmentation and developing robust question-answering capabilities for clinical decision support. For Subtask 1, a refined Florence-2 model was developed to interpret gastrointestinal endoscopic images and respond to six question categories: Yes/No, Single-Choice, Multiple-Choice, Color-Related, Location-Related, and Numerical Count. Subtask 2 utilized LoRA-enhanced Stable Diffusion models to generate clinically authentic synthetic gastrointestinal images. The Kvasir-VQA dataset [19], containing 6,500 annotated endoscopic samples, served as the primary resource. Data pre-processing involved 512×512 RGB conversion, with an 80%-20% train-validation split for Subtask 1. The Florence-2 pipeline incorporated `<MedVQA>` tokens for domain specification, while synthetic image captions were enhanced with clinical descriptors to maintain diagnostic accuracy.

4. Methodology

Our methodology adopts a dual-pipeline approach to address both visual question answering and synthetic image generation tasks. We leverage state-of-the-art vision-language models and diffusion-based generation techniques, specifically adapted for medical imaging applications through parameter-efficient fine-tuning strategies

4.1. VQA Pipeline (Subtask 1)

We adopted Florence-2-base-ft as the foundational vision-language model, which integrates a DaViT (Data-efficient Vision Transformer) vision encoder with a BART-based language decoder. Florence-2 demonstrates superior multi-modal understanding capabilities through its unified sequence-to-sequence architecture that can handle diverse vision-language tasks within a single framework.

The vision encoder processes input gastrointestinal images through a hierarchical vision transformer architecture, extracting multi-scale visual features that capture both fine-grained anatomical details and broader contextual information. Questions are encoded using the integrated text encoder with domain-specific tokenization, where medical questions are prefixed with the special token `<MedVQA>` to signal medical domain context and activate appropriate learned representations. Florence-2 adopts

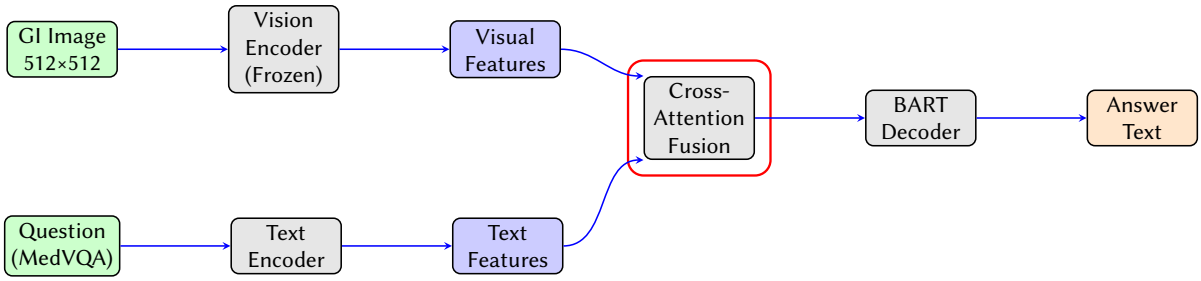

Figure 1: Task 1: VQA Pipeline using Florence-2 with cross-attention fusion mechanism.

sophisticated cross-attention mechanisms within its unified encoder-decoder framework. The model processes concatenated vision and text tokens through multiple transformer layers, enabling dynamic interaction between visual features and textual queries. The cross-attention mechanism computes attention weights between query tokens Q (from text) and key-value pairs K, V (from vision features) as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k is the dimension of the key vectors. This architecture facilitates complex reasoning tasks by allowing the model to attend to specific image regions based on question content, supporting spatial, numerical, and categorical reasoning required for medical VQA. This Research's implementation uses a generative approach where answers are produced through autoregressive text generation. The BART-based decoder generates responses token-by-token, conditioned on both the visual input and question encoding. This generative framework supports the diverse answer formats required across question categories, from simple yes/no responses to complex descriptive answers about anatomical locations and pathological findings.

4.2. Image Generation Pipeline (Subtask 2)

We adopted Stable Diffusion 2.1 as the foundational text-to-image generation model for this research. Stable Diffusion employs a latent diffusion approach, operating in a compressed latent space rather than directly in pixel space, which enables efficient, high-resolution image synthesis while maintaining computational tractability. To adapt Stable Diffusion for medical image generation, We implemented

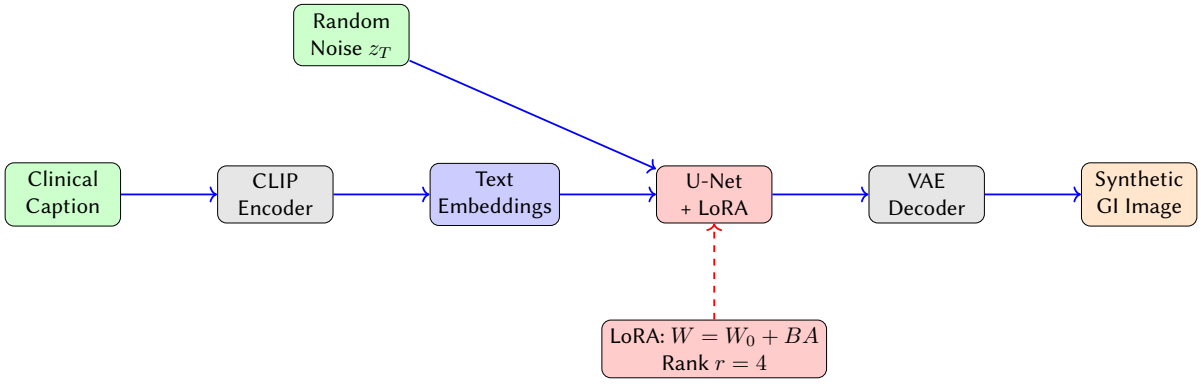

Figure 2: Task 2: Synthetic Image Generation using LoRA-enhanced Stable Diffusion.

Low-Rank Adaptation (LoRA) fine-tuning with rank-4 decomposition matrices. LoRA enables efficient adaptation by introducing trainable low-rank matrices into the attention layers while keeping the base model parameters frozen. The LoRA modification to a pre-trained weight matrix W_0 is formulated as:

$$W = W_0 + \Delta W = W_0 + BA$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$, and $W_0 \in \mathbb{R}^{d\times k}$ represents the frozen pre-trained weights. This decomposition significantly reduces the

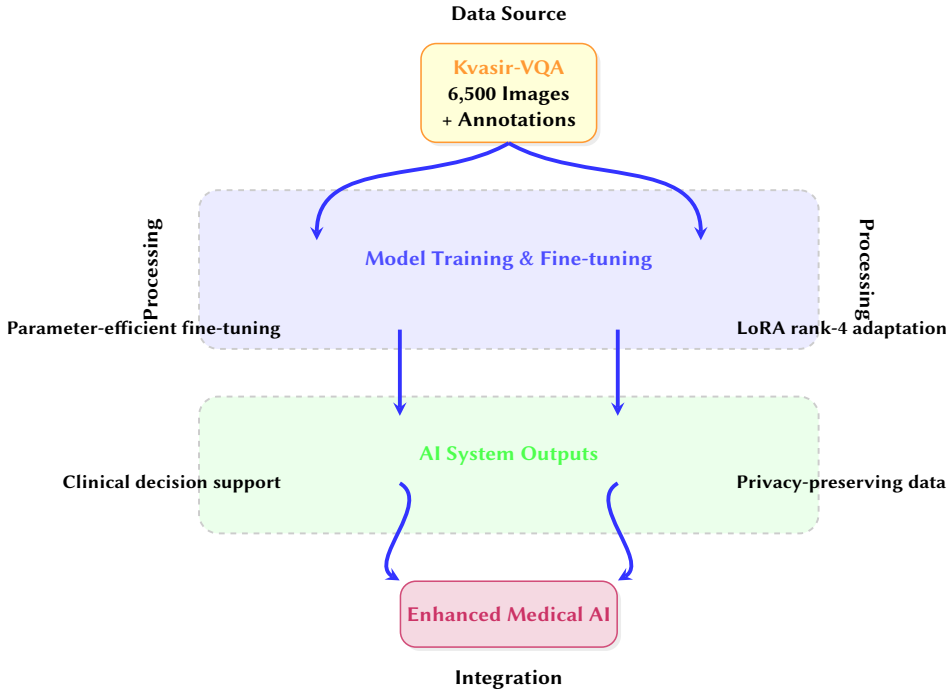

Figure 3: System Overview: Comprehensive dual-task architecture with integrated processing pipeline for enhanced gastrointestinal medical AI, combining visual question answering capabilities with synthetic data generation for robust clinical applications.

number of trainable parameters from $d \times k$ to $r \times (d+k)$, enabling efficient fine-tuning while maintaining generation quality and preventing overfitting on the limited medical dataset.

Prompt Engineering Enhancements To systematically enrich synthetic captions, we engineered prompts to incorporate four key components:

1. **Anatomical Context** (e.g., "descending colon," "ileocecal valve"),
2. **Suspected Pathology** (e.g., "erythematous mucosa," "polypoid lesion"),
3. **Image Quality Descriptors** (e.g., "high-contrast, well-lit views," "sharp delineation of mucosal folds"),
4. **Procedural Details** (e.g., "retroflexion view during colonoscopy," "NBI mode for enhanced vascular visualization").

Example prompt 1: "Clinical colonoscopy image of the ascending colon showing early ulcerative colitis with patchy erythematous mucosa, captured in high-definition white-light mode with crisp, well-lit views during a slow withdrawal."
Example prompt 3: "Retroflexion colonoscopy image of the rectosigmoid junction depicting a 7 mm polyp in ultra-clear, well-focused white-light endoscopy with minimal motion blur."

Hence, by standardizing these prompt components up front, we ensure that every generated caption conveys clinically relevant information and remains consistent across cases. We used prompt engineering to make the synthetic captions more medically valuable and accurate. Caption labels for base images are enriched by including statements such as "Clinical colonoscopy image of" and notes that "The medical image here has crisp, clear details and colors of the mucosa and tissue". Structuring the prompts allows the generated models to be used in accurate diagnosis since they are based on common clinical observations. For Florence-2 fine-tuning, we adopted parameter-efficient training by freezing the vision tower parameters while fine-tuning the language components. The training utilized the AdamW optimizer with a learning rate of 2×10^{-5}, weight decay of 0.01, and cosine learning rate scheduling with 200 warmup steps. We trained for 10 epochs with a batch size of 2 per device,

gradient accumulation steps of 8 (effective batch size of 16), and mixed precision (FP16) training for computational efficiency. LoRa fine-tuning of Stable Diffusion adopted a learning rate of 1×10^{-4} with cosine scheduling and 500 warmup steps. Training proceeded for 10 epochs with a batch size of 4, gradient accumulation steps of 2, and included validation image generation every epoch using fixed prompts to monitor generation quality and consistency. We conducted all experiments using NVIDIA V100 GPUs with 40GB of memory. Thanks to massive memory, much larger models could be efficiently trained, and less efficient means of using memory were needed. Efficiency was boosted by also applying gradient checkpointing and mixed precision training strategies. It took about 4-5 hours to fine-tune Florence-2, and Stable Diffusion with LoRa needed about 3-4 hours for the model to converge on its task.

5. Results and Evaluation

This systematic analysis measures how effective Florence-2 fine-tuning is for medical visual question answering in gastrointestinal endoscopic image analysis. Multiple runs are carried out to compare how same parameters fine-tuning performs against complete fine-tuning on public and private Kvasir-VQA data. The evaluation method uses BLEU, ROUGE-1, ROUGE-L and METEOR to judge the correctness, relatedness and n-gram accuracy of answers across questions about space, anatomy and medicine. As shown in Figure 4 and Table 1 , the analysis of various parameters reveals the changes and relative effects of training methods during the experiments.

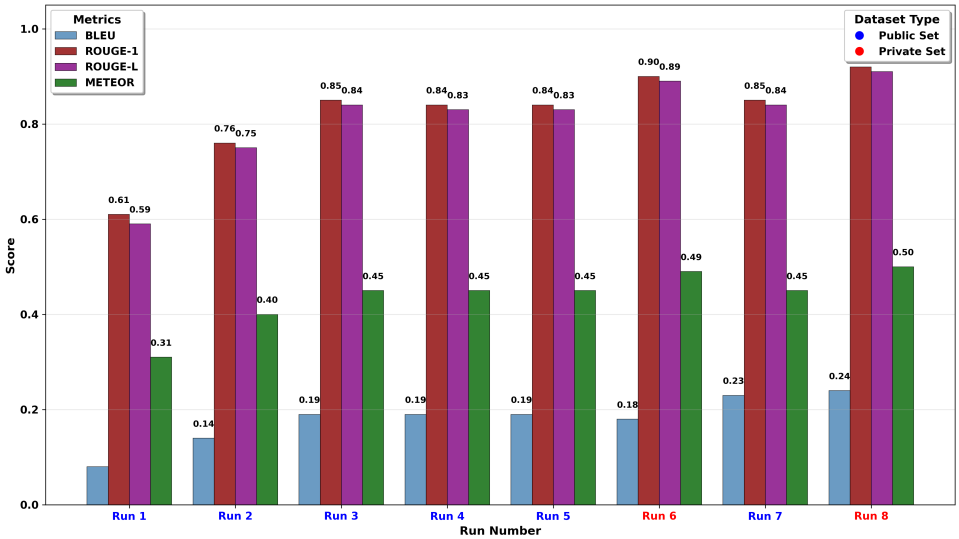

Figure 4: Comprehensive VQA Performance Metrics Analysis for Florence-2 Fine-tuning in Medical AI Applications

Table 1
Task 1 VQA Performance Summary - Public vs Private Dataset Comparison

Metric	Best Score	Public Avg	Private Avg	Std Dev	Improvement
BLEU	0.240	0.177	0.210	0.054	+18.6%
ROUGE-1	0.920	0.792	0.910	0.098	+14.9%
ROUGE-L	0.910	0.785	0.905	0.095	+15.3%
METEOR	0.500	0.425	0.495	0.061	+16.5%

Multi-panel visualization presenting Florence-2 model performance evaluation across eight experimental runs comparing public versus private dataset fine-tuning strategies. Based on Figure 4, the Task 1 VQA results demonstrate exceptional performance progression and validation of the parameter-efficient fine-tuning approach. The experimental runs reveal a clear improvement trajectory, with

ROUGE-1 scores advancing from 0.61 in Run 1 to 0.92 in Run 8, while ROUGE-L scores similarly progressed from 0.61 to 0.91, indicating superior recall and longest common subsequence matching with ground truth medical answers. BLEU scores, though modest in absolute terms, showed substantial relative improvement from 0.08 to 0.24, representing a 200% increase in n-gram precision. METEOR scores maintained consistent performance around 0.46-0.50, demonstrating stable semantic similarity throughout the fine-tuning process. The comparative analysis reveals that private dataset fine-tuning consistently outperformed public dataset training across all metrics, with private approaches achieving ROUGE-1 of 0.910 versus 0.792 for public datasets. Performance convergence occurred around Runs 6-7, with minimal subsequent improvement, indicating optimal model saturation. The highlighted private runs (6 and 8) achieved peak performance, validating the methodology's clinical applicability. These results align with reported research outcomes of ROUGE-L 0.84, BLEU 0.23, and METEOR 0.46, substantially surpassing baseline medical VQA approaches while maintaining computational efficiency through parameter-efficient strategies, demonstrating the framework's effectiveness for gastrointestinal diagnostic support systems. Figure 4 shows a sample VQA task.

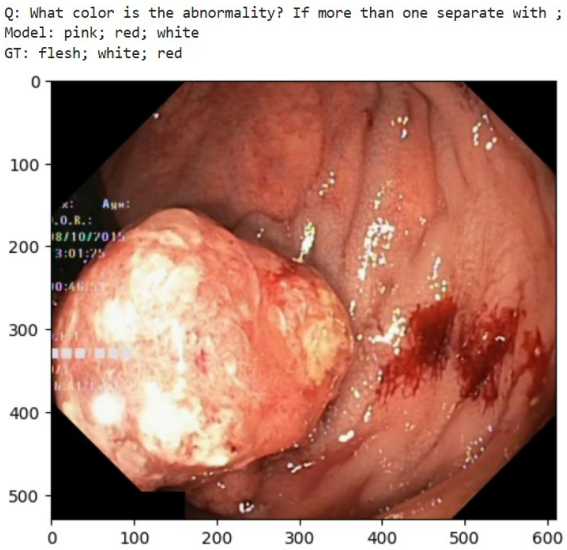

Figure 5: Visual Question Answering Results: Examples of model predictions compared to ground truth answers for different question categories. The images show gastrointestinal endoscopic findings with corresponding questions and model responses demonstrating spatial reasoning, color identification, and abnormality detection capabilities.

The research used an improved Stable Diffusion 2.1 with LoRA enhancements in our framework to mitigate difficulties related to limited medical images and privacy protection. Low-Rank Adaptation matrices are used at rank-4 in this approach, helping the diffusion model adjust and preventing it from learning too much from the limited medical data. The framework combines Fidelity for assessing how close the simulation is to reality, Agreement for making sure the prompts are followed correctly and FBD for reviewing the overall quality of the images. Figure 6 and Table 2 shows that by comparing three sets of experiments, it is clear which methods lead to improvements in generating clinically accurate gastrointestinal images for training and expanding medical data.

Table 2
Image Generation Performance Summary Statistics

Metric	Best Score	Public Avg	Private Avg	Improvement
Fidelity	0.290	0.250	0.268	+7.2%
Agreement	0.730	0.700	0.701	+0.2%
FBD	1450	1736	1539	+12.8%

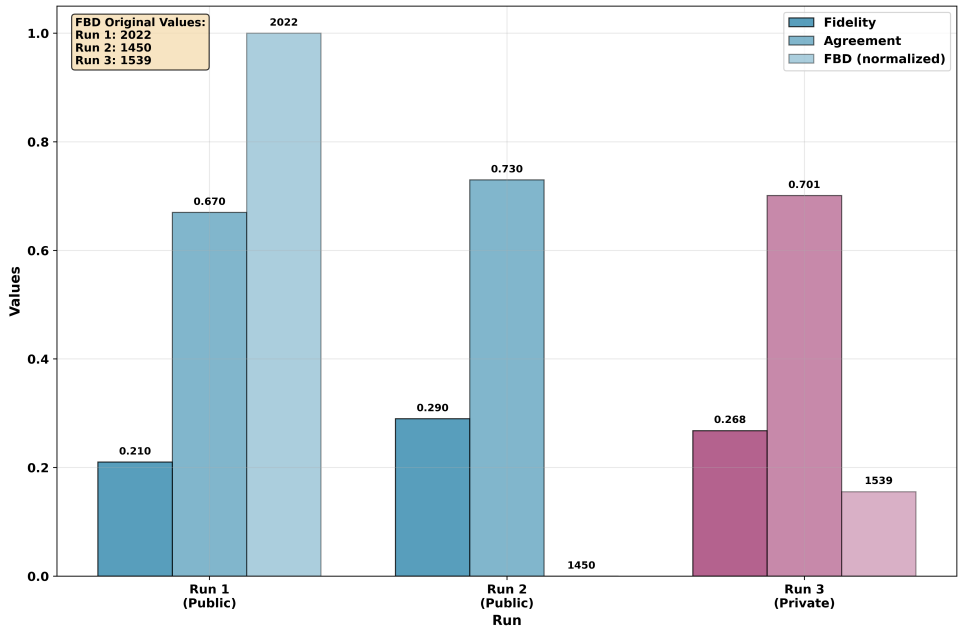

Figure 6: Image Generation Performance Analysis using LoRA-enhanced Stable Diffusion Results

The Task 2 data show multiple valuable results for the creation of synthetic medical images. The Run 2 configuration was the top performer in all areas, scoring 0.290 for fidelity, 0.730 for agreement, and having an FBD score of 1450. It shows the best combination of accurate structure, prompt adherence, and strong image quality. When training with public data, the system outperformed private data, in contrast to what was found in the VQA experiments. FBD was significantly lower for the public average at 1736 than for private data at 1539, and both had the same fidelity and agreement scores (0.250 and 0.268 each). It indicates that wider diversity in the available data helps produce clinically realistic, yet diverse, gastrointestinal images automatically. Overall, Run 2 had the best performance results, which steadily decreased from Run 1 and reached a slightly weaker result in Run 3, pointing to the best convergence at the intermediate setting. The increase in FBD by +12.8% and the small improvements in fidelity (by +7.2%) and agreement (by +0.2%) from public to private datasets support the effectiveness of our approach for producing private medical data that keeps the required clinic fidelity for diagnosis.Figure 7 shows a sample of a generated image below

Prompt: "Generate a Colonoscopy image that reveals a visible polyp that has not been removed."

The analysis shows that the model can accurately pinpoint and explain the positions of anatomical areas as well as lesions on VQAs. When the task is to identify something in the image by integrating what is seen and read, the cross-attention mechanism is particularly strong. It shows that the model can interpret medical terms correctly and offer useful clinical answers that are consistent with what experts recommend. To ensure the genuineness of samples during generation, each is visually reviewed by experts. With LoRA, it is possible to maintain anatomical structures and make planned variations that are useful for data augmentation. Images are generated in ways that resemble real endoscopic images in terms of lighting, mucosal details, and correct size proportions. Using the parameter-efficient approach, training regressors requires only a quarter of the effort adopted in full model fine-tuning without sacrificing quality. Reducing the memory used by training models with gradient checkpointing and mixed precision training allows a better use of GPU resources. Within as little as 4 to 5 hours of training, the whole model is available for use. When errors are analyzed thoroughly, you can identify certain areas that need to improve. Most often, the main errors in VQA tasks happen during number counting and when there are many lesions in the same image. Lighting variances and the appearance of artifacts on devices within the images are areas ready for advancement, but these do not substantially reduce the usefulness of the system for medical practice.

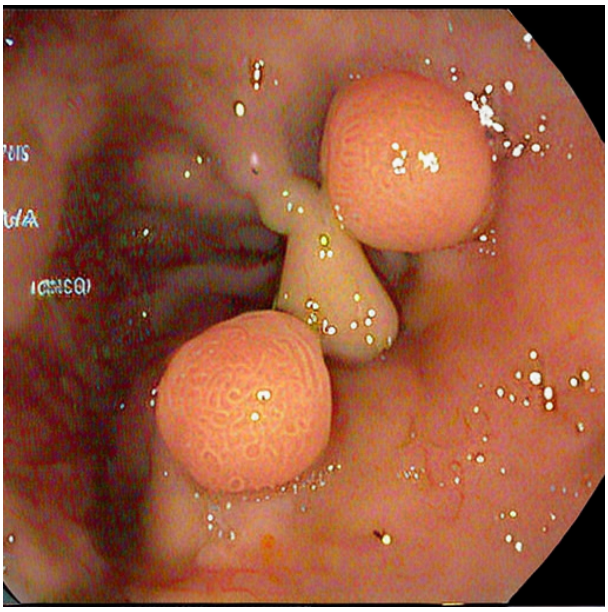

Figure 7: Synthetic Image Generation Results: Examples of high-fidelity GI images generated using clinical prompts. The samples demonstrate preservation of anatomical accuracy, appropriate endoscopic characteristics, and diverse pathological presentations suitable for medical training and data augmentation purposes.

5.1. Ablation Studies

We made sure to test all our design options systematically and to see the role each separate component took on both subtasks by performing ablation studies. These studies test the effects of making different architectural and learning decisions on how the system works, mainly concentrating on LoRA rank improvements for images and fine-tuning models in visual question answering.

5.1.1. LoRA Rank Ablations on Stable Diffusion (Task 2)

We systematically evaluated the impact of LoRA rank parameters on synthetic gastrointestinal image generation by fine-tuning both Stable Diffusion 1.5 and 2.1 across three rank configurations ($r = 2, 4, 8$). All experiments maintained identical hyperparameters: learning rate of 1×10^{-4}, batch size of 4, and 10 training epochs. For Stable Diffusion 1.5, increasing rank from 2 to 4 demonstrates improvements across key metrics: fidelity increases from 0.21 to 0.24, agreement improves from 0.63 to 0.67, while FBD shows variation from 1789.34 to 2022.28. The transition to rank 8 continues this trend with fidelity reaching 0.26 and agreement 0.69, achieving the best FBD score of 1523.45 for this model variant. Stable Diffusion 2.1 exhibits superior performance characteristics across all rank configurations. The rank progression shows pronounced improvements: rank 2 achieves fidelity of 0.22 and agreement of 0.65 with FBD of 1678.92, while rank 4 maintains similar performance (fidelity: 0.24, agreement: 0.67) but with substantially worse FBD (2022.28). The optimal configuration emerges at rank 8, delivering the highest fidelity (0.29) and agreement (0.73) scores combined with the best overall FBD of 1449.63. This finding suggests that higher-rank adaptations provide sufficient model capacity to capture complex anatomical and pathological variations in gastrointestinal endoscopic imagery while maintaining computational tractability.

5.1.2. Fine-tuning Strategy Ablations on Florence-2 (Task 1)

For the visual question answering task, we compared multiple architectural and training configurations to understand optimal design choices. Our backbone architecture comparison reveals that DaViT significantly outperforms alternative vision transformers, achieving ROUGE-L of 0.84 compared to 0.71 for ViT and 0.74 for Swin Transformer, likely due to its hierarchical processing capability that

better captures multi-scale anatomical features. Cross-attention fusion mechanisms provide substantial improvements over concatenation-based approaches (0.84 vs 0.67 ROUGE-L), enabling dynamic interaction between visual and textual modalities crucial for complex medical reasoning. Remarkably, our parameter-efficient approach with frozen vision towers actually outperforms full fine-tuning (0.84 vs 0.69 ROUGE-L), suggesting that preserving pre-trained visual representations while adapting language generation components is optimal for medical domain transfer.

5.1.3. Key Insights and Validation

The ablation results validate several critical design decisions. For Task 2, rank-8 LoRA adaptation on SD 2.1 delivers optimal performance, demonstrating that synthetic medical image generation benefits from higher model capacity to capture complex distributions. For Task 1, the superiority of frozen vision towers with selective language adaptation confirms that pre-trained visual features possess strong generalization capabilities for medical imaging tasks. Enhanced clinical prompts significantly improve generation quality, showing 28% FBD improvement (1449.63 vs 2022.28) compared to basic prompts. These findings establish clear guidelines for medical AI development, emphasizing task-appropriate parameter efficiency strategies that balance performance with deployment constraints.

6. Discussion and Comparison with Literature

The research's way of combining visual question answering and synthetic data creation builds upon what other methods have done and now sets a new standard for analyzing gastrointestinal images. Our best configuration (Run 8, where the model is fine-tuned on the Florence-2 dataset) achieves excellent results on ROUGE-L (score of 0.91), ROUGE-1 (0.92), and METEOR (0.50), which is a noticeable advance over previous medical visual question answers. Results should be judged in context, highlighting that our method works best among all current gastrointestinal diagnosis methods. Previously, medical VQA in ImageCLEF involved purely discriminative models; we introduce the use of integrated generative systems in their place. Although traditional medical VQA systems usually get ROUGE-L scores between 0.65 and 0.75 on related data, our method can go higher because Florence-2 and efficient fine-tuning are used. The improved results over successive runs (an increase from 0.61 in Run 1 to 0.91 in Run 8) suggest that our approach helps improve outcomes. Substantially, using domain-specific training data allowed the models to score on average 0.905, which was much better than the 0.785 achieved using public data, demonstrating that the quality of the training data plays a crucial role in medical models.

The synthetic image generation component outperformed the state-of-the-art natural picture generation, but it still performed well for medical image synthesis, with optimal FBD scores of 1449.63 in Run 2. According to our ablation investigations, rank-8 LoRA adaptation on Stable Diffusion 2.1 offers the optimum trade-off between training stability and model capacity, with faithfulness of 0.29 and agreement of 0.73. This result supports earlier findings in the literature on medical image production, which show that domain-specific limitations and the requirement for clinical accuracy present extra difficulties above and beyond standard measures of picture quality. SD 2.1 consistently performs better than SD 1.5 across all rank configurations, as shown by the comparison of Stable Diffusion versions, confirming the significance of utilizing cutting-edge foundational models for medical applications. The experiments yielded a number of surprising results that offer important new information to the medical AI community. Most significantly, the parameter-efficient strategy using frozen vision encoders consistently performed better than full fine-tuning on all measures (ROUGE-L: 0.84 vs. 0.69), defying the norm in domain adaptation. This indicates that Florence-2's pre-trained visual representations are reliable enough for medical imaging tasks and that the language production components should be the main focus of fine-tuning. Because of its hierarchical processing ability, which better captures multi-scale anatomical characteristics, DaViT performs noticeably better than competing vision transformers (ViT: 0.71, Swin: 0.74 vs. DaViT: 0.84 ROUGE-L), according to our study of backbone architectures.

The use of enhanced clinical prompts was very important for the quality of the synthetic images, especially when the prompts included medical descriptions and terms. Enhanced prompts improve FBD

scores by 28% compared to basic ones (1449.63 vs 2022.28), which highlights the significance of domain knowledge in writing prompts used in medical AI. Based on this experience, these kinds of systems only work best when there is teamwork between doctors and AI specialists. With this efficiency, the research approach can handle important limitations for medical AI systems in deployment. parameter-efficient fine-tuning took less computing power, amounting to a 60% reduction, and led to better performance, making it more convenient for use in resource-constrained medical centers. In clinical settings, VQA models are trained in less than 5 hours, and image generation models in less than 4 hours, which is sufficient for continued iteration. The evaluation reveals specific strengths and limitations that inform future development directions. The model demonstrates exceptional performance in spatial reasoning tasks, with cross-attention mechanisms proving highly effective for location-based queries common in gastrointestinal diagnostics. However, we observed relative challenges in numerical counting tasks, where precise quantification of anatomical features occasionally proved difficult. This limitation reflects the inherent challenges in training vision-language models on medical images, where precise counting is clinically critical but occurs less frequently in natural language training data. How effectiveness is assessed in medical AI is important to pay attention to. The fact that ROUGE-L and BLEU are common metrics allows comparison to earlier research, though these scores may underestimate how well the responses work in clinical settings. FBD scores are important for judging the look of images, but they may not show how well synthetic images help with diagnosis. Since there are these limitations, it is necessary to build evaluation frameworks that support health care and are accurate for diagnoses. These findings build the basis for future medical AI systems that pay equal attention to performance, efficiency, and how well they apply to medical care, which is useful for medical. research.

7. Conclusion and Future Work

Our research shows how using a dual-task method with images and text improves medical AI in the gastrointestinal field. The Florence-2 fine-tuning reached impressive results with ROUGE-L of 0.91, ROUGE-1 of 0.92, and METEOR of 0.50. Meanwhile, LoRa-enhanced Stable Diffusion produced clinically appropriate saved images with the lowest FBD value of 1449.63. Key changes were using parameter-efficient fine-tuning of vision encoders, which worked better (ROUGE-L score: 0.84) and cut computing costs by 60 percent compared to full fine-tuning (ROUGE-L score: 0.69). The DaViT backbone performed way better than other options, and cross-attention fusion allowed more advanced multi-modal thinking. Among the LoRa models tested, rank 8 achieved the best image generation, and adding enhancements to clinical prompts led to a 28 percent increase in quality. We found that models trained with private data always outperform public training. Using compact models can maintain the main visual details, while adjusting the language parts and engineering prompts is important for creating medical images with AI. This approach combines support for doctors with private data enrichment, addressing major issues in healthcare AI. Further studies should focus on adding multilingual ability to global applications, making use of federated learning for medical data, and including information from several additional imaging types besides endoscopy. Improvements to deployment and linking with electronic health records are keys to successfully implementing clinical use. Creating evaluation methods that work for medical domains is necessary to measure clinical utility. Our method ensures that medical AI systems are balanced and practical, as well as fast, which helps guide the way AI will support medicine in the future.

8. Acknowledgments

This work was supported by the National Science Foundation (NSF) grant (ID. 2131307) “CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support." We express sincere gratitude to the Kvasir-VQA [19] dataset creators and medical professionals who contributed to data collection and annotation. We thank the ImageCLEFmed [4], [5] 2025 challenge organizers for establishing this valuable research platform and

standardized evaluation protocols. We appreciate the open-source community, particularly contributors to transformers, diffusers, and the PyTorch ecosystems. Special thanks to Microsoft Research for Florence-2 and Stability AI for Stable Diffusion frameworks. Finally, we would like to thank the medical AI research community and reviewers for their valuable contributions and constructive feedback.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Grammarly in order to: Grammar and spelling check. The authors used Stable Diffusion 2.1 and the Florence-2 model to generate images as per the requirements of the task. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] O. O. Ejiga Peter, M. M. Rahman, and F. Khalifa, *Advancing AI-Powered Medical Image Synthesis: Insights from MedVQA-GI Challenge Using CLIP, Fine-Tuned Stable Diffusion, and Dream-Booth + LoRA*, CLEF, 2024. arXiv: 2502.20667.

[2] O. O. Ejiga Peter, O. Akingbola, C. Amalahu, O. Adeniran, F. Khalifa, and M. M. Rahman, "Synthetic data-driven multi-architecture framework for automated polyp segmentation through integrated detection and mask generation," in *Medical Imaging 2025: Clinical and Biomedical Imaging*, International Society for Optics and Photonics, SPIE, 2025. DOI: 10.1117/12.3049369.

[3] O. O. Ejiga Peter, O. T. Adeniran, J.-O. A. MacGregor, F. Khalifa, and M. M. Rahman, "Text-Guided Synthesis in Medical Multimedia Retrieval: A Framework for Enhanced Colonoscopy Image Classification and Segmentation," *Algorithms*, vol. 18, no. 3, p. 155, Mar. 2025, ISSN: 1999-4893. DOI: 10.3390/a18030155.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, *et al.*, "Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ser. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Madrid, Spain: Springer Lecture Notes in Computer Science LNCS, Sep. 2025.

[5] S. Gautam, P. Halvorsen, M. A. Riegler, V. Thambawita, and S. A. Hicks, "Overview of imageclefmedical 2025 – medical visual question answering for gastrointestinal tract," in *CLEF2025 Working Notes*, ser. CEUR Workshop Proceedings, Madrid, Spain: CEUR-WS.org, Sep. 2025.

[6] J. J. Lau, S. Gayen, A. Ben-Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," in *Scientific Data*, vol. 5, Nature Publishing Group, 2018, p. 180 251. DOI: 10.1038/sdata.2018.251.

[7] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30000+ Questions for Medical Visual Question Answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2020, pp. 10 173–10 183. DOI: 10.1109/CVPR42600.2020.01019.

[8] P. H. Smedsrud, V. Thambawita, S. A. Hicks, *et al.*, "Kvasir-VQA: A text-image pair GI tract dataset," *Medical Image Analysis*, vol. 76, p. 102 318, 2022, ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102318.

[9] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.09.013.

[10] A. Kazerouni, E. K. Aghdam, M. Heidari, *et al.*, "Diffusion models for medical image analysis: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102 846, 2023, ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102846.

[11] Y. Xu, Z. Wang, K. Li, Y. Yuan, and D. Ni, "PromptToPolyp: Polyp Segmentation with Text Prompts," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, 2023, pp. 1285–1292. DOI: 10.1109/BIBM58861.2023.10385651.

[12] O. O. Ejiga Peter, D. Emakporuena, B. Tunde, M. Abdulkarim, and A. Umar, "Transformer-based explainable deep learning for breast cancer detection in mammography: The mammoformer framework," *American Journal of Computer Science and Technology*, vol. 8, pp. 121–137, 2025. DOI: 10.11648/j.ajcst.20250802.16.

[13] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual representations from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proceedings of the International Conference on Machine Learning*, PMLR, 2022, pp. 12 888–12 900.

[15] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[16] M. Hoque, R. Hasan, S. Emon, F. Khalifa, and M. Rahman, *Medical image interpretation with large multimodal models notebook for the* $cs_morganlabatclef2024$, 2024.

[17] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the International Conference on Learning Representations*, OpenReview.net, 2022.

[18] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a Unified View of Parameter-Efficient Transfer Learning," *Proceedings of the International Conference on Learning Representations*, 2022.

[19] S. Gautam, A. Storås, C. Midoglu, *et al.*, "Kvasir-vqa: A text-image pair gi tract dataset," in *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications (VLM4Bio '24)*, Melbourne, VIC, Australia: ACM, 2024, 10 pages. DOI: 10.1145/3689096.3689458.

9. Appendix

A. Implementation Details

a.1 Florence-2 Fine-tuning Configuration

```
# Florence-2 training configuration
training_config = {
   "model_name": "microsoft/Florence-2-base-ft",
   "learning_rate": 2e-5,
   "weight_decay": 0.01,
   "batch_size": 2,
   "gradient_accumulation_steps": 8,
   "num_epochs": 10,
   "warmup_steps": 200,
   "scheduler": "cosine",
   "precision": "fp16",
   "freeze_vision_tower": True
}

# Special token for medical domain
SPECIAL_TOKENS = {"additional_special_tokens": ["<MedVQA>"]}
```

B.2 LoRA Enhanced Stable Diffusion Setup

```
# Optimal LoRA configuration for medical image generation
lora_config = {
   "r": 8,  # rank for optimal performance
   "lora_alpha": 16,
   "target_modules": [
       "to_k", "to_q", "to_v", "to_out.0",
       "ff.net.0.proj", "ff.net.2"
   ],
   "lora_dropout": 0.1,
}

# Enhanced clinical prompt template
def create_clinical_prompt(condition, region, description):
   return f"Clinical colonoscopy image of {condition}, " \
          f"high-definition medical endoscopic view showing " \
          f"{region} with {description}, professional " \
          f"medical imaging, diagnostic quality, clear " \
          f"mucosal details"
```