International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

HYBRID SYSTEM FOR VISUAL QUESTION LOCATION GENERATION AND ANSWERING IN GASTRO-INTESTINAL ENDOSCOPY IMAGES

Rajeswari Jayaraman 1*[0000-0003-0705-0548], Kavitha Srinivasan1[0000-0003-3439-2383],, Divyasri Krishnakumar1[0009-0004-6081-3027], Cyril Melvin Vincent 1[0009-0007-7173-086X].

[1,*] Department of Computer Science Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai, 603110, Tamil Nadu, India.

[*]Corresponding author(s). E-mail(s): rajeswarij@ssn.edu.in

Contributing authors: (kavithas@ssn.edu.in; divyasri2011037@ssn.edu.in; cyrilmelvin2010009@ssn.edu.in)

All authors contributed equally to this work.

Abstract

The Visual Question Location Generation and Answering (VQLGA) is a hybrid system, which combines the Visual Question Answering (VQA) and Visual Location-based Question Answering (VLQA) approaches together for assisting the healthcare professionals. VQLGA system focuses on gastrointestinal (GI) endoscopy images of ImageCLEF med 2023 challenge, detecting and preventing colorectal cancer. The GI endoscopy images generated through endoscopy procedures are complex in nature, therefore analyzing these images are time consuming for medical experts. To address this challenge a hybrid system is designed and implemented using suitable Deep Learning techniques with appropriate quantitative metrics for validation at each stage of process. VQLGA system, leverages VGG16 for image features and LSTM for text features in the VQA model achieving an accuracy of 80%. The VLQA model, driven by DenseNet121 and UNet architectures, attains a remarkable Intersection over Union (IoU) score of 83.4%. By integrating these models, the hybrid VQLGA system achieved an accuracy of 71% and IoU of 81.8% is evidence for enhanced diagnostic efficiency through proposed system. Additionally, the segmented output of the VQLGA system is validated using Explainable AI (XAI) technique to support the segmented region through visualization.

Math. Subj. Classification 2020:68T07 ,68U10,68T50,68T07, 68U10.

Keywords: ImageCLEF, Visual Question Answering, Visual Location based Question Answering, Explainable AI, Colonoscopy images, Polyps.

1. Introduction

Gastro-Intestinal (GI) images acquired through endoscopy are essential for the early detection and prevention of colorectal cancer, allowing medical professionals to identify polyps, lesions, and other abnormalities in the colon. However, interpreting these images is complex and time-consuming due to factors such as variability in appearance, overlapping or hidden lesions, low image quality, and the need for precise differentiation between benign and malignant growths. Additionally, real-time analysis requires expertise to ensure accurate

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

diagnosis and effective patient management, making the process both challenging and critical in colorectal cancer prevention and treatment. To address these challenges a novel hybrid system: Visual Question Location Generation and Answering (VQLGA) is proposed. This system combines Visual Question Answering (VQA) and Visual Location-based Question Answering (VLQA) approaches to automate the analysis and interpretation of colonoscopy images.

Visual Question Answering (VQA) is a model that generates textual responses to image-based queries by extracting features from both images and text. It employs VGG16 for image feature extraction and LSTM for textual processing, enabling the identification of abnormalities, handling multi-answer questions, and responding to open-ended medical queries, particularly in colorectal cancer diagnosis. However, traditional VQA lacks precise localization, which limits its interpretability in medical imaging. To overcome this limitation, Visual-Question Location-based Answering (VLQA) is introduced, focusing on location-aware feature extraction to enhance the accuracy and justification of responses [6]. VLQA employs advanced deep-learning techniques such as DenseNet121, which enhances feature representation by retaining critical image details, and UNet segmentation, which enables precise localization of abnormal regions like polyps and tumors [7]. By combining VQA's text-based responses with VLQA's location-aware analysis, the advanced hybrid approach can significantly enhance the colorectal cancer detection, improving both accuracy and interpretability in medical imaging. Because the generated answers are not only textually relevant but also visually justified, addressing key limitations of traditional AI models in medical imaging. However, as AI adoption in healthcare grows, the need for explainable AI (XAI) becomes increasingly crucial to ensure transparency, trust, and regulatory compliance. Conventional deep-learning models often act as "black boxes," providing little insight into how decisions are made, which is a significant concern in critical applications like cancer diagnosis.

The VQA-VLQA based hybrid approach with XAI improves both textual and visual explanations, making AI decisions more interpretable for clinicians. The remaining sections of this research paper are organized as: recent work, methodology, analysis and results, finally ended with conclusion and its future scope.

2. Recent works

This section provides an overview of the ImageCLEF Med tasks for medical images, followed by an exploration of the 2023 task related to gastrointestinal (GI) endoscopy images for medical VQA and VQLG with its merits and demerits.

Medical Visual Question Answering (VQA) has observed significant advancements across various ImageCLEF Medical VQA tasks, with an increasing shift from predefined models to transformer-based models for language and image encoding. In ImageCLEF 2018 VQA task, the various models explored are image encoders like VGG16, ResNet152, and Inception ResNet v2 pre-trained on the ImageNet dataset, along with language encoders such as LSTM and BiLSTM. These models are fused to generate VQA predictions, achieved BLEU scores ranging from 0.054 to 0.188, with attention-based fusion performing better than simple concatenation [4,23,24,26-29]. Interestingly, in the research paper [25] leveraged GRU for

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

language encoding and element-wise multiplication for fusion, outperforming models that relied on attention networks for concatenation. In ImageCLEF 2019, research shifted towards transformer-based models like BERT, which improved accuracy, reaching 0.938 compared to 0.789 achieved by LSTM-based models [1,9].

In ImageCLEF 2023, researchers explored various approaches to improve medical VQA performance. Vision Transformer, Visual BERT, and SegFormer were integrated, with SegFormer achieving the highest accuracy of 95.7% for binary questions [15]. Other methods fused BioBERT, LSTM, and VGG16 using the Stacked Attention Network (SAN) technique [12,17,18] while a top-scoring model employed BEiT for image feature extraction and ALBERT for text feature extraction [10,19]. Additionally, the wsq4747 team introduced a novel architecture combining BLIP-2, ViT-G, and GLM-6B, refining GLM-6B through a two-stage process, which improved image feature extraction and achieved 0.7396 accuracy in polyp predictions [22]. For the VLQA task, authors of paper [10] trained two separate models for polyps and instruments due to dataset constraints. They implemented a semantic segmentation model using the Detectron2 library and applied transfer learning by fine-tuning a Mask R-CNN model previously trained on the COCO dataset, achieving an accuracy of 99% and 94%, respectively [16].

The literature [21] compares EfficientDet, Faster R-CNN, YOLOv3, YOLOv4, and DeepLabV3+ with ColSegNet, which achieved a higher Dice score of 88.72. Paper [5] benchmarks endoluminal scene segmentation in colonoscopy, addressing polyp miss rates and malignancy assessment, with FCN8 achieving an IoU of 59.5% [3]. Papers [5,11] introduce DRi-Net and Graft-U-Net, modified UNet models with wider networks and deeper encoder-decoder layers, improving IoU to 81.38%-86%, 2% higher than UNet [8]. Paper [14] applies Multi-modal Multi-task Learning (M3L) for classification before CNN-based segmentation [13]. Paper [23] explores CNN with Grad-CAM, using heatmaps for segmentation, showing partial tumor localization but suggesting that combining Grad-CAM with other methods could improve accuracy [2].

The literature review highlights key gaps in medical Visual Question Answering (VQA) systems, including limitations in pretrained transformers, difficulty in handling open-ended and multi-answer questions [20], lack of explainability, and challenges in combining perception with reasoning. Current segmentation models also lack in detecting small polyps and validating the identified segments. To address these issues, the research aims to develop a robust VQA and VLQA approaches that can manage complex questions, improve segmentation accuracy, integrate perception and reasoning, and provide confidence levels for abnormality detection. Then the combined VQA-VLQA system is proposed to enhance both text-based answers and visual segmentations. Additionally, explainability techniques like XAI Grad-CAM is used to verify the predicted outcome for improving the trust and reliability in medical AI.

3. Methodology

In the proposed research VQA, VLQA and a hybrid system (VQLGA) are designed and implemented using one of the opensource dataset of ImageCLEF 2023. In VQA model, an input image is given through VGG16, and a question is provided to predict the corresponding

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

answer. In a VLQA model, an input image with a mask is fed into the system, using UNet and DenseNet for object detection. The system is trained to predict the mask, identifying features like polyps and instruments. The proposed hybrid VQLGA system, the outputs from VQA (predicted answer) and VLQA (predicted mask), along with the original image and question, are fed into the model. The VQLGA system produces a textual answer and, if applicable, displays a segmented image. This segmented image is validated using XAI Grad-CAM, which generates a heatmap to verify the accuracy of the segment as shown in the Figure 1.

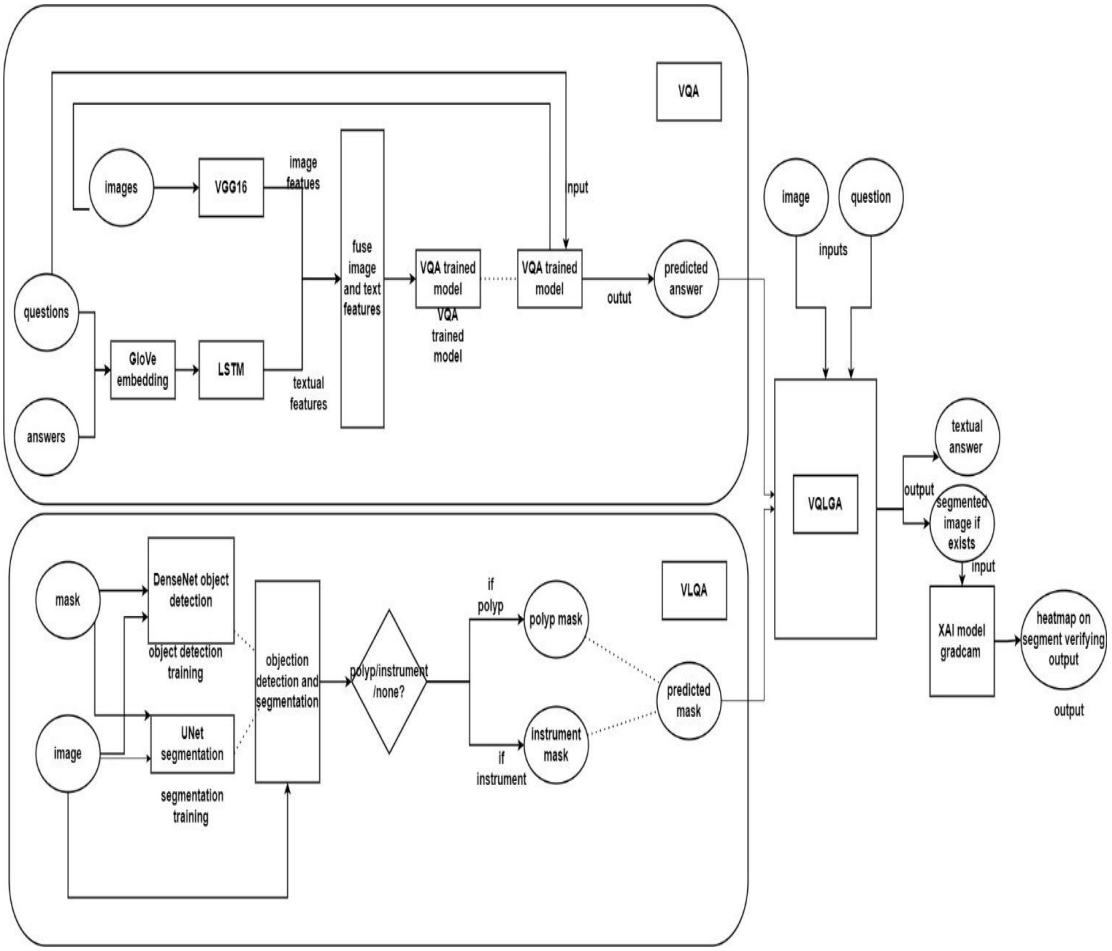

Figure 1 - System design of proposed VQLGA

3.1 Dataset description

The dataset of VQA contains 1998 images with 20 questions each, totaling 39,960 samples. However, only 18 questions with textual responses are considered, and irrelevant questions are marked as "Not Relevant" and two more open-ended questions are added to the dataset such as 'Tell me about the polyps' and 'Tell me about the instruments'.

The dataset is preprocessed by addressing inconsistencies, removing invalid samples, and converting the JSON file into a data frame with comma-separated multiple answers. It is then converted to lowercase, and approximately 70-30 ratio of train-test split is performed, ensuring proportional representation from each question type. As a result, the QA pairs of training set contains 27,960 and the test set contains 12,000, totally 39,960 as given in Table 1.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In VLQA, approximately 682 images are given with masks, out of which 499 belong to polyps and 183 to instruments. The images are rotated by 45 degrees, flipped horizontally and vertically along with its mask which brings the count to 1996 and 732 respectively during augmentation. The images are split in the 80-20 ratio for train and test and the train images are further split into 80-20 ratio for train and validation as given in Table 2.

Table 1

VQA: Training and test set details of images with QA

Dataset	**Input**		**Output (Answer)**
	Images	**Questions**	
Training set	1,998	27,960	27,960
Test set	1,998	12,000	12,000
Total	**3,996**	**39,960**	**39,960**

Table 2

VLQA: Training, test and validation set split of augmented data

Question of mask type image	**Count**	**Augmented data**	**Training set**	**Test set**	**Validation set**
Polyp	500-1 = 499	499*4 = 1996	1276	400	320
Instruments	183	183*4 = 732	468	147	117

3. 2 Visual Question Location Generation and Answering (VQLGA)

Visual Question Location Generation and Answering (VQLGA) system is a combination of VQA and VLQA. Given an image and its associated question, the VQLGA model returns a textual answer and a predicted mask for the image. The segment accompanying the textual answer provides a visual guide for diagnosticians and non-medical personnel to trust the textual answer as a location segment accompanies it. The same image and question are passed to two models such as VQA and VLQA, where one model returns a textual answer, and the other model returns a segment. The segment is further passed to a GradCAM XAI model as shown in Figure 2.

Image and text feature extraction and fusion: The images are resized into 224 X 224 pixels for height and width and the image features are extracted using VGG16. The text features, questions and answers are converted to lower case and tokenized. Further converted to numerical sequences and padded to the maximum length of the question as vectors using GloVe vectorization technique. The answers are converted to one Hot encoded labels. The image

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

features and textual features from the question are multiplied, normalized, and passed through a Dense layer.

Train the Model: The fused features are used to train the model for 15 epochs with a learning rate of 0.001. Adam optimizer is used along with a loss function of categorical cross entropy since it is a multiclass classification problem. The accuracy is saturated at 79% after 10 epochs and the run was stopped at that point.

Test the Model: The model is tested on the 30 percent of the data which is 12,000 samples. The number of questions of each type in the test dataset is about 600.

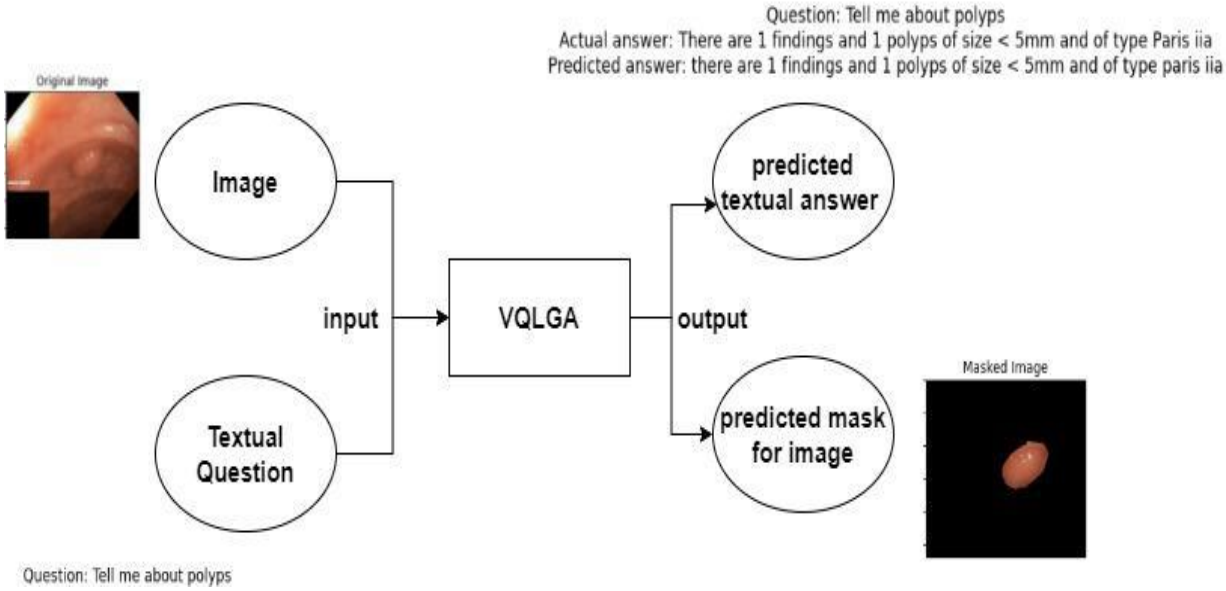

Figure 2 – System design of VQLGA

3.3 Visual Location Question Answering (VLQA)

Object detection model: This model is used to identify the whether the image has a polyp or instrument using an object detection algorithm (DenseNet121). Since this model is pre-trained on the ImageNet dataset, its weights are predefined, and it only needs to be fine-tuned using the polyp and instrument images as per dataset split given in Table 2.

Training the model: Two different models are considered for segmentation - UNet and SegNet. Of those, UNet provided the best results during a trial run, and hence, it was applied for the final model training as shown in the Figure 3. The model uses Adam optimizer and binary cross-entropy loss function. The model was trained for 25 epochs on a batch size of 16. The validation accuracy was saturated at 0.9626 for instruments after 7 epochs, with a validation loss of 0.1014. The validation accuracy was saturated at 0.8300 at 10 epochs for polyps, with a validation loss of 0.3756.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Testing the model: The trained UNet segmentation model gives a mean IoU of 0.802. The segmentation model performed better on the polyps. For instruments, the IoU was 0.77. For predicting the mask, it takes approximately 67 milliseconds. The threshold is set at 0.5 for the mask generation.

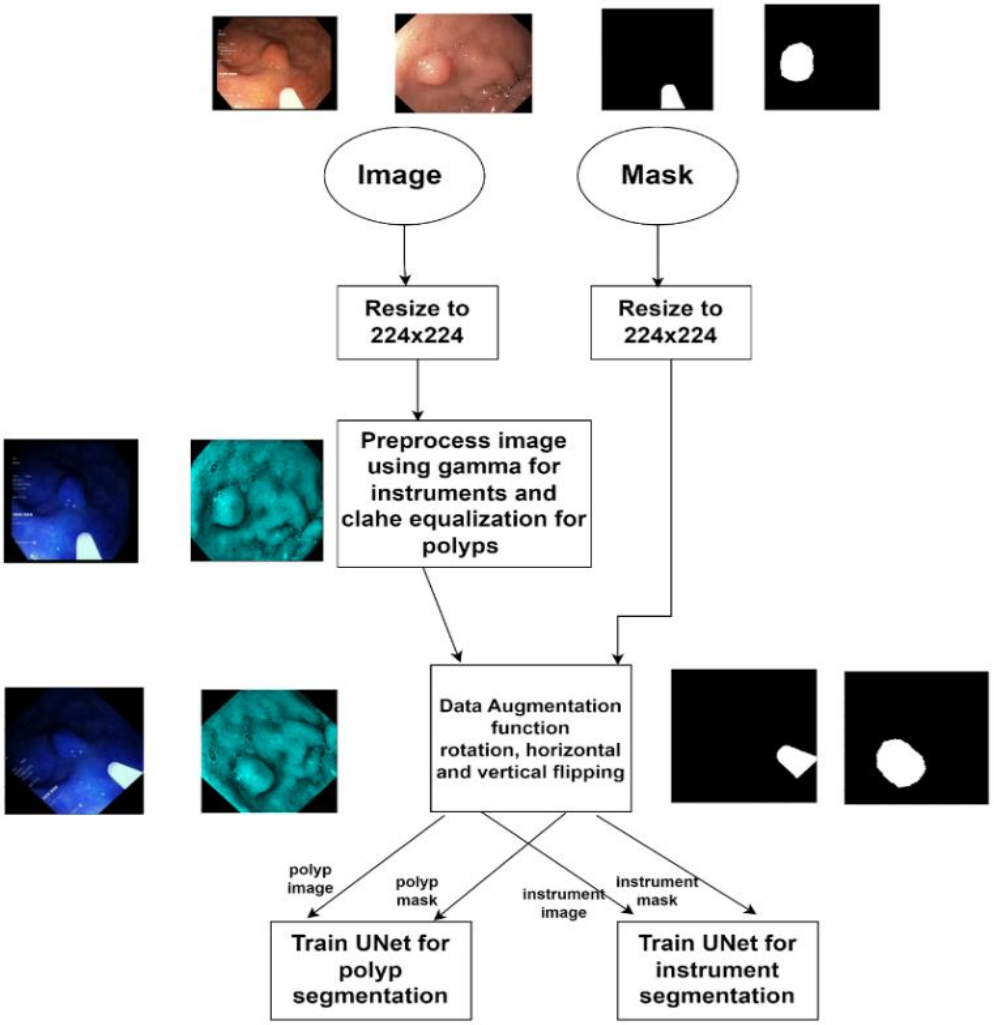

Figure 3 - System design of VLQA

Algorithm: *Training VLQA Model and Segment Prediction*

Input: *datapts* including *Image* (*Img*), its corresponding object labels (*labels*), and its corresponding Mask (*mask*).

test data refers to the Images for which output needs to be generated.

Output: The mask predicted by the model for the given image is returned. The mask is superimposed on the image to get the segment of the image.

cnf level refers to the confidence level with which a feature is present in the image.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1: **if** *Img* in *invalid datapts* **then**
2: Remove *Img* from *data pts*
3: **end if**
4: **for** *Img* in *test data* **do**
5: *cnf level,label* ← Object _detection model(*Img*(*i*))
6: **if** *cnf level* > 0.85 and *label* is 'polyp' **then**
7: *Img segment* ← Trained UNetseg polyp(*Img*)
8: **else if** *cnf level* > 0.85 and *label* is 'inst' **then**
9: *Img segment* ← Trained UNetseg inst(*Img*)
10: **else**
11: **return** 'No feature is found in the image'
12: **end if**
13: **return** *Img segment*
14: **end for**=0

3.4 Explainable AI

Explainable AI (XAI) is crucial for transparency in complex models like those used in medical image processing. One popular technique, Grad-CAM, highlights important regions in an image for the model's prediction as shown in the Figure 4. In medical segmentation, it helps to verify results by showing which areas contributed most to the decision. For instance, in colonoscopy polyp segmentation, after the AI segments polyps, Grad-CAM can visually highlight significant regions for verification against ground truth. This overlay of Grad-CAM on the segmented polyps creates a visual tool for validation and understanding model behavior, fostering trust and collaboration between AI and medical professionals.

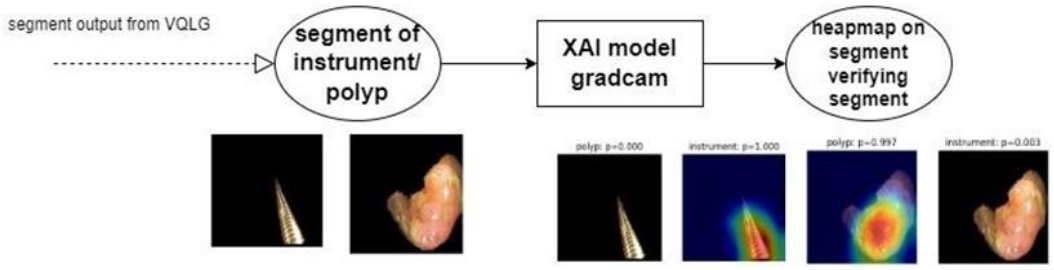

Figure 4 : Sample XAI Output

3.4.1 GRAD-CAM: Gradient-weighted class activation mapping

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Gradient Computation: The gradient of the output class score with respect to the feature maps of the last convolutional layer captures the importance of each feature map for the target class prediction which is polyp or instrument here in the equation 1.

$$\frac{\partial y^C}{\partial A_k} \tag{1}$$

Importance weights: computed for each feature map by applying global average pooling to the gradients. These weights signify the relevance of each feature map in predicting whether the feature maps to a polyp or an instrument are shown in below equation 2.

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^C}{\partial A_k}(i,j) \tag{2}$$

Class Activation Map (CAM) Generation: The CAM is generated by linearly combining the weighted feature maps followed by a Rectified Linear Unit (ReLU) activation in equation 3. This process highlights regions within the feature maps that significantly contribute to the target class prediction.

$$L_{ACM} = (\sum \alpha_k A_k) \tag{3}$$

Localization Visualization: The CAM is resized to the size of the input image using bilinear interpolation to obtain the Grad-CAM. This visualization technique provides insights into the spatial locations in the input image that are crucial for the model's classification decision as shown in equation 4.

$$Grad - CAM(x,y) = \sum \alpha_k A_k\ (x,y) \tag{4}$$

Where:

A_K denotes the k-th feature map of the last convolutional layer.

Y^c represents the output class score before softmax for the target class c.

(i, j) represents the spatial location within the feature maps.

Z is a normalization factor.

ReLU denotes the Rectified Linear Unit function.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. Experimental results

In this section the metrices used in the analysis VQA, VQLA and the proposed VQLGA are given and the values obtained are analyzed and validated.

4.1 Quantitative metrics

The metrics related to classification and segmentation (VQA,VLQA) are listed below:

Accuracy: Accuracy refers to the percentage of correct predictions made by a generated model as given in Equation 5. The formula for calculating accuracy involves considering the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) of the predictions of a given query.

$$Accuracy = \frac{Number\ of\ Correctly\ classified\ smaples}{Total\ number\ of\ samples} \tag{5}$$

Intersection over Union (IoU): IoU is used to evaluate the performance of segmented region using object detection algorithms. It quantifies the overlap between the predicted bounding box or segmented region and the ground truth bounding box. The formula for IoU is given in Equation 6. A higher IoU value indicates a better alignment between the predicted and actual regions.

$$IOU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{6}$$

BLEU score: This score is computed based on the generated answer and reference answers. The formula is given in Equation 7.

$$BLEU = BP.exp\ exp\left(\sum_{n=1}^{N} \omega_n\ log\ log\ p_n\right) \tag{7}$$

Where BP is the Brevity Penalty, which penalizes short, generated answers, P_n is the modified precision for n-grams of length, W_n is the weight for the n-gram precision and N is the maximum order of n-gram.

4.2 Performance analysis of VQA and VLQA

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In VQLA the segmentation of region is carried out using SegNet and UNet, From the analysis, we inferred that UNet took lesser training time almost 50% lesser than SegNet and achieved a better IoU are shown in the Tables 3 and 4. In Table 3, the accuracy of each question for the test set is listed with its train and test size. In Table 4, IoU score of polyp and instrument is mentioned with its segmentation model and accuracy. From the results, it is observed that the segmentation on polyps is better than segmentation on instruments. This could be because of the number of training instances available, there are only 183 images of instruments and its masks for training while there are 499 polyps available for training.

Table 3

Accuracy of individual QA with its Train and Test split in VQA

Question	**Accuracy (%)**	**Training set**	**Test set**
are there any abnormalities in the image?	94.27	1422	576
are there any anatomical landmarks in the image?	92.15	1361	637
are there any instruments in the image?	89.93	1442	556
have all polyps been removed?	95.83	1422	576
how many findings are present?	79.41	1420	578
how many instruments are in the image?	91.58	1392	606
how many polyps are in the image?	94.87	1394	604
is there a green/black box artefact?	62.59	1426	572
is there text?	80.95	1389	609
is this finding easy to detect?	77.80	1408	590
tell me about instruments	72.56	1406	592
tell me about polyps	63.51	1406	592
what color is the abnormality?	54.52	1378	620
what color is the anatomical landmark?	94.55	1411	587
what is the size of the polyp?	74.87	1401	597
what type of polyp is present?	79.81	1374	624
what type of procedure is the image taken from?	98.73	1370	628
where in the image is the abnormality?	63.17	1398	600
where in the image is the anatomical landmark?	72.06	1386	612

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 4

Performance Analysis of UNet based Image Segmentation

Section	Model	Accuracy	IOU
VQA (standard dataset)	VGG16	71.1%	–
VLQA (polyp)	UNet	–	81.8%
VLQA (instrument)	UNet	–	75.5%

Tables 5 and 6 infers the performance of VQA and VLQA models and analyzes the impact of data augmentation. VGG16 performs better than ResNet50 in VQA, reaching 95% accuracy for binary questions. In VLQA, UNet proves effective, achieving 83.4% IoU for polyp detection. Data augmentation enhances segmentation, with 45-degree rotation improving instrument IoU to 72.8% and 90-degree rotation boosting polyp IoU to 72.4%, These results highlight the importance of model selection and augmentation technique in enhancing VQA performance and segmentation accuracy.

Table 5

Results of VQA and VQLA

Section	Model	Accuracy	BLEU	IOU
VQA (standard dataset)	ResNet50	53%	–	–
VQA (binary questions)	VGG16	95%	–	–
VQA (standard dataset)	VGG16	80%	–	–
VQA (open ended)	VGG16	–	0.59	–
VLQA (polyp)	UNet	–	–	83.4%
VLQA (instrument)	UNet	–	–	77%

Table 6

Results of VLQA after augmentation of data

Data Augmentation Technique	Instrument Mask IoU (%)	Polyp Mask IoU (%)
None	54.569	62.097
45-degree Rotation	72.8	69.92
90-degree Rotation	66.30	72.4
Horizontal Flip	68.12	70.05

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Vertical Flip	66.03	68.98

4.3 Performance analysis of VQLGA

A higher IoU obtained from the object detection implies the close matching between the predicted and actual regions and the accuracy indicates the reliable predictions in colonoscopy images, justifies the improvement in the overall model performance. IoU is higher for the combined model (VQLGA) as mentioned in Table 7. It is noticeable that segmentation of polyps faired better than instruments as per IoU, since the dataset size is bigger.

Table 7

Polyp mask and Instrument type images Vs Evaluation metrics for VQLGA

Image	**IoU**	**Dice**	**Precision**	**Recall**	**F1**
Polyp	83.4%	0.869	0.956	0.94989	0.974306
Instrument	77%	0.860	0.989	0.95081	0.9747899

5. Conclusion and future works

In this research work the images of colonoscopy with its relevant QA pairs and its masks for two questions on polyp and instrument are considered in developing three types of models. The models are Visual Question Answering (VQA), Visual Location Question Answering (VLQA) along with XAI and Visual Question Location Generation and Answering (VQLGA) as a hybrid approach of VQA and VLQA. The formulation of open-ended questions and the generation of multiple answers for a single question is a novel venture in medical VQA. Moreover, in VLQA generated segment and its predicted text answer are justified with the use of GradCAM XAI by highlighting the pertinent features in an image for a given question. The proposed VQLGA model resulted a F1 score of 97% for both polyp and instrument type images whereas for IoU metric it is higher for polyp images with 83.4%. To improve the model's performance further, a large dataset is recommended. Future research should focus on reducing RAM usage for real-time diagnostics. Incorporating real-time feedback from medical professionals can enhance accuracy and training with diverse data types and creating detailed reports will help doctors in real time.

References

1. A.Lubna, S. Kalady, and A. Lijiya. *MoBVQA: A Modality based Medical Image Visual Question Answering System*, TENCON 2019–2019 IEEE Region 10 Conference, Kochi, India, pp. 727–732 (2019).
2. Chin Yii Eu, Tong Boon Tang, ChengHung Lin, Lok Hua Lee and Cheng-Kai Lu. *Automatic Polyp Segmentation in Colonoscopy Images Using a Modified Deep Convolutional Encoder-Decoder Architecture* , in Sensors, Vol. 21, No. 16, pp. 5630. doi: 10.3390/s21165630 (2021).

Received: August 11, 2025

3. D.Jha et al. *Real-Time PolypDetection, Localization and Segmentation in Colonoscopy Using Deep Learning* , IEEE Access, Vol. 9, pp. 40496-40510. doi: 10.1109/ACCESS.2021.3063716 (2021).
4. Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, and Anh Nguyen. *Multiple Meta-model Quantifying for Medical Visual Question Answering*, In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, Proc.Part V Springer-Verlag, Berlin, Heidelberg,pp. 64–74. doi: 10.1007/978-3-030-87240-37 (2021).
5. D. Vazquez, A. M. Lopez, et al. *A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images* , in Hindawi Journal of Healthcare Engineering, Vol.2017, Article ID 4037190, 9 pages. doi:10.1155/2017/4037190 (2017).
6. H. Borgli, et al. *HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy*, Scientific Data, Vol. 7, No. 1, 283. doi: 10.1038/s41597-020-00622-y (2020).
7. H. M. Afify, K. K. Mohammed, and A. E. Hassanien. *An improved framework for polyp image segmentation based on SegNet architecture* , International Journal of Imaging Systems and Technology, Vol. 31, No. 3, pp. 1741-1751. doi: 10.1002/ima.22568 (2021).
8. M. Ramzan, M. Raza, M.I. Sharif, and S.Kadry. *Gastrointestinal Tract Polyp Anomaly Segmentation on Colonoscopy Images Using Graft-U-Net* , Journal of Personalized Medicine, Vol. 12, No. 1459. doi:10.3390/jpm12091459 (2022).
9. Noor Mohamed, Sheerin Sitara, and Srinivasan Kavitha. *A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges* , Journal of Intelligent & Fuzzy Systems, Vol. 44, No. 4, pp.5803-5819. doi: 10.3233/JIFS-222569 (2023).
10. Patrycja Cieplicka, Julia Klos, Maciej Morawski, and Jaroslaw Opala. *Language-based Colonoscopy Image Analysis with Pretrained Neural Networks* , CLEF2023 Working Notes, Proceedings of the 14th International Conference of the CLEF Association, No. 120. doi: 10.48550/arXiv.2307.02783 (2023).
11. P. Narayan Sholapur, I. M. *Explainable AI and Deep Learning techniques for Colon Cancer Detection*, 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 1096-1105. doi:10.1109/ICAC3N49924.2022.9850145 (2022).
12. Rohit Raj Gunti and Abebe Rorissa. *A Dual of Stacked Attention Networks (SAN's) and VGG-16 Model-Based Visual Question Answering Evaluation Working notes for the ImageCLEFmed MEDVQA-GI Lab at CLEF 2023* , Proc. CEUR Workshop, Thessaloniki, Greece, pp. 1–12 (2023).
13. Azizi, F. Mustafa, J. Kim, E. Horng, R. Liaw, R. Pfeffer, and D. K. Miller. *Multi-Modal Multi-Task Learning for Medical Image Classification* , IEEE Journal of Biomedical and Health Informatics, Vol. 24, No. 3, pp. 819-829. doi:10.1109/JBHI.2019.2944157 (2020).

Received: August 11, 2025

14. Shuyue Guan and Murray Loew. *A Sneak Attack on Segmentation of Medical Images Using Deep Neural Network Classifiers* , pp 1-8 (2022).

15. S. S. Noor Mohamed, K. Srinivasan, and R. Gopalsamy. *SSN MLRG at MEDVQA-GI 2023: Visual Question Generation and Answering using Transformer based Pre-trained Models* , in Proc of the 14th International Conference of the CLEF Association (CLEF 2023), Thessaloniki, Greece (2023).

16. Steven Hicks, Andrea Storas, Pal Halvorsen,Thomas de Lange, Michael Riegler, and Vajira Thambawita. *Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract* , CLEF2023 Working Notes, Proceedings of the 14th International Conference of the CLEF Association, No.107. doi:10.48550/arXiv.2307.02783 (2023).

17. S. Upadhyay and S. S. Tripathy. *BITMesra at ImageCLEF 2023: Fusion of Blended Image and Text Features for Medical VQA* ,CEUR Workshop Proceedings, pp. 1-12 (2023).

18. S. Wang, W. Zhou, Y. Yang, H. Huang, Z.Ye, T. Zhang, and D. Yang. *Adapting Pre-Trained Visual and Language Models for Medical Image Question Answering Notebook for the Baidu Intelligent Health Unit and Peng Cheng Laboratory Joint Team at CLEF 2023* , in Proc of the Conference and Labs of the Evaluation Forum (CLEF), Thessaloniki, Greece (2023).

19. Tascon-Morales, S., Marquez-Neila, P., and Sznitman, R. *Consistency-Preserving Visual Question Answering in Medical Imaging* , In: Wang, L., Dou, Q., Fletcher, P.T.,Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention MICCAI. Lecture Notes in Computer Science, vol 13438. doi: 10.1007/978-3-031-16452-137 (2022)

20. T.Van Sonsbeek, A. Radford, C. Olah, and L. A. Hendricks. *Medical Visual Question Answering Through Prefix Tuning of Language Models* , arXiv preprint arXiv:2303.05977 (2023).

21. X.Lan, H. Chen, and W. Jin. *DRINet: segmentation of polyp in colonoscopy images using dense residual-inception network* , Front. Physiol., Vol. 14, pp. 1290820 (2023).

22. Y.Bazi, F. Al-Hamadi, M. A. Mahmoud, I.A. F. Al-Sharif, P. M. Atkinson, and H. A. El-Zaart. *Vision–Language Model for Visual Question Answering in Medical Imagery* , Bioengineering, Vol. 10, No. 3, pp.380. doi: 10.3390/bioengineering1003038 (2023).

23. Zhou, Y., Kang, X., Ren, F., *Employing Inception-Resnetv2 and Bi-LSTM for medical domain visual question answering, in: CLEF* (Working Notes) (2018).

24. Peng, Y., Liu, F., Rosen, M.P., *UMass at ImageCLEF medical visual question answering (Med-VQA) 2018 task*, in :CLEF (Working notes) (2018).

25. Ambati, R., Reddy Dudyala, C., *A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering, in: 2018* , 15th IEEE India Council International Conference (INDICON), pp. 1–6. doi:10.1109/INDICON45594.2018.8987108 (2018).

26. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., *Demner-Fushman, D., 2018. NLM at ImageCLEF 2018 visual question answering in the medical domain*., in: CLEF (Working Notes) (2018).

Received: August 11, 2025

27. Gupta, D., Suman, S., Ekbal, A., *Hierarchical deep multimodal network for medical visual question answering,* Expert Systems with Applications 164, 113993, doi.org/10.1016/j.eswa.2020.113993 (2021).
28. Allaouzi, I., Ahmed, M.B., *Deep neural networks and decision tree classifier for visual question answering in the medical domain*., in: CLEF (Working Notes) (2018).
29. Talafha B , B., Al-Ayyoub, M., *JUST at VQA-Med: A VGGSeq2Seq model*, in: CLEF (Working Notes) (2018).

Rajeswari Jayaraman, working in the Department of Computer Science and Engineering, Srisivasubaramaniya Nadar College of Engineering, Chennai, India. Research domains encompass Computer Vision, Data Analytics, AI, and Machine Learning, rajeswarij@ssn.edu.in.

Kavitha Srinivasan, working in the Department of Computer Science and Engineering, Srisivasubaramaniya Nadar College of Engineering, Chennai, India. Area of Research encompass Machine Learning, Medical Image Processing, Explainable AI Computer Vision, kavithas@ssn.edu.in.

Divyasri Krishnakumar, doing Masters in Uinversity of Tornoto , area of intrest , Machine Learning, Medical Image Processing, Natural Language Processing, Explainable AI, divyasri2011037@ssn.edu.in.

Cyril Melvin Vincent , doing Master's in North Carolina State University, Boston, Massachusetts, United States. Research domains encompass Software Engineering, Algorithms, Neural Networks, Software Security, Game Engine Foundations, cyrilmelvin2010009@ssn.edu.in