Visual Question Answering (with Multimodal Explanations) for Gastrointestinal Imaging

Prabavathy Balasundaram* Ganga B† Suguneswari K‡ Pradeep M§

Department of Computer Science,
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, Tamil Nadu, India

Abstract

Visual Question Answering (VQA) in the gastrointestinal (GI) imaging domain supports AI-assisted endoscopic diagnosis, as highlighted in the Medico 2025 Challenge (Subtask 1). This paper presents an **Ultra-Efficient VQA Model**, designed as a lightweight early-fusion framework for efficient and accurate medical image understanding. The model integrates a ResNet-18 visual encoder with a Bidirectional LSTM text encoder, followed by a shallow Multi-Layer Perceptron for answer prediction. Grad-CAM is employed to provide visual explanations, ensuring model decisions align with clinically relevant regions. Evaluated on the Kvasir-VQA-x1 dataset, the system achieves strong accuracy and METEOR scores while maintaining low computational cost. The results establish a practical baseline for explainable multimodal GI-VQA systems, facilitating reproducibility and supporting the development of AI-driven clinical decision support tools.

1 Introduction

Gastrointestinal (GI) diseases, particularly colorectal cancer (CRC), pose significant global health challenges. Early detection through endoscopic imaging is essential but often limited by human interpretation, which can be time-consuming and inconsistent. Artificial Intelligence (AI) has emerged as a valuable aid in medical imaging, offering the potential to enhance diagnostic accuracy and consistency.

Visual Question Answering (VQA) combines computer vision and natural language processing to provide meaningful answers to questions based on images. In the medical context, VQA enables AI systems to address clinically relevant queries, improving interpretability and clinician trust. The *Medico 2025 Challenge*, specifically Subtask 1, employs the Kvasir-VQA-x1 dataset containing annotated endoscopic images, serving as a strong foundation for benchmarking multimodal models in GI imaging.

This paper presents the **Ultra-Efficient VQA Model**, an early-fusion approach integrating a lightweight ResNet-18 CNN for visual feature extraction and a Bi-LSTM for textual understanding. The

*Email: prabavathyb@ssn.edu.in
†Email: ganga2520068@ssn.edu.in
‡Email: suguneswari2520028@ssn.edu.in
§Email: pradeep2520042@ssn.edu.in

model emphasizes both efficiency and interpretability, using Grad-CAM to visualize attention regions. This work establishes a reproducible baseline for GI VQA, demonstrating that reliable and explainable performance can be achieved without computationally expensive transformer-based models.

Contributions

The main contributions of this study include:

1. Development of a compact and efficient VQA architecture using CNN and Bi-LSTM for GI image analysis.
2. A reproducible performance baseline for the Kvasir-VQA-x1 dataset in the Medico 2025 Challenge.
3. Integration of explainability via Grad-CAM for enhanced clinical interpretability.

2 Related Works

2.1 General Visual Question Answering (VQA) Architectures

Visual Question Answering (VQA) combines visual and textual reasoning to answer questions about images. Early models such as the Stacked Attention Network (SAN) [1] and Multimodal Compact Bilinear Pooling (MCB) [2] improved multimodal fusion through attention mechanisms. Transformer-based architectures like LXMERT [3], ViLBERT [4], and VisualBERT [5] further enhanced fine-grained image–text interactions via cross-modal attention. ViLT [6] simplified the pipeline using Vision Transformers. Despite strong results, these models are computationally heavy, motivating lightweight CNN–LSTM frameworks for efficient inference.

2.2 Medical Visual Question Answering

Medical VQA extends general VQA to clinical imagery. The VQA-Med challenges [7] introduced benchmark datasets connecting medical images and expert-generated questions. Later datasets like PathVQA [8] focused on pathology reasoning. The Kvasir-VQA-x1 dataset [10] for gastrointestinal (GI) imaging covers binary, counting, and reasoning question types, providing a diverse benchmark for evaluating multimodal medical AI systems.

2.3 Explainability and Multimodal Fusion in Medical AI

Interpretability remains essential in clinical AI. Explainable AI (XAI) techniques such as Grad-CAM [9] visualize regions influencing model predictions. Transformer-based medical VQA systems and multimodal fusion frameworks emphasize explainable models for clinically trustworthy predictions [10].

3 Proposed Methodology

3.1 Dataset and Preprocessing

This study employs the **Kvasir-VQA-x1** dataset, comprising 159,549 question–answer (Q–A) pairs derived from 6,500 gastrointestinal (GI) endoscopic images [10]. The data were split into training, validation, and testing sets in a 70:15:15 ratio.

Preprocessing involved cleaning and tokenizing the questions by removing punctuation, converting text to lowercase, and padding to a maximum of 50 tokens. Images were resized to 224×224 pixels, normalized, and augmented with random flips and rotations to enhance generalization.

3.2 Model Architecture

The proposed **Ultra-Efficient VQA Model** adopts an early-fusion framework integrating visual and textual information:

- **Visual encoder:** A pretrained ResNet-18 extracts 512-dimensional visual embeddings.
- **Text encoder:** A Bi-directional LSTM processes tokenized questions to capture contextual meaning.
- **Fusion and prediction:** The concatenated image–text features are fed to a two-layer MLP with ReLU activation and softmax output for answer prediction.

3.3 Training Strategy

The model was optimized using the Adam optimizer (learning rate = 0.001) and cross-entropy loss. A batch size of 64, dropout (0.5), and early stopping (patience = 5) were used to prevent overfitting. Model performance was evaluated using Accuracy, BLEU, METEOR, and ROUGE metrics.

3.4 Explainability

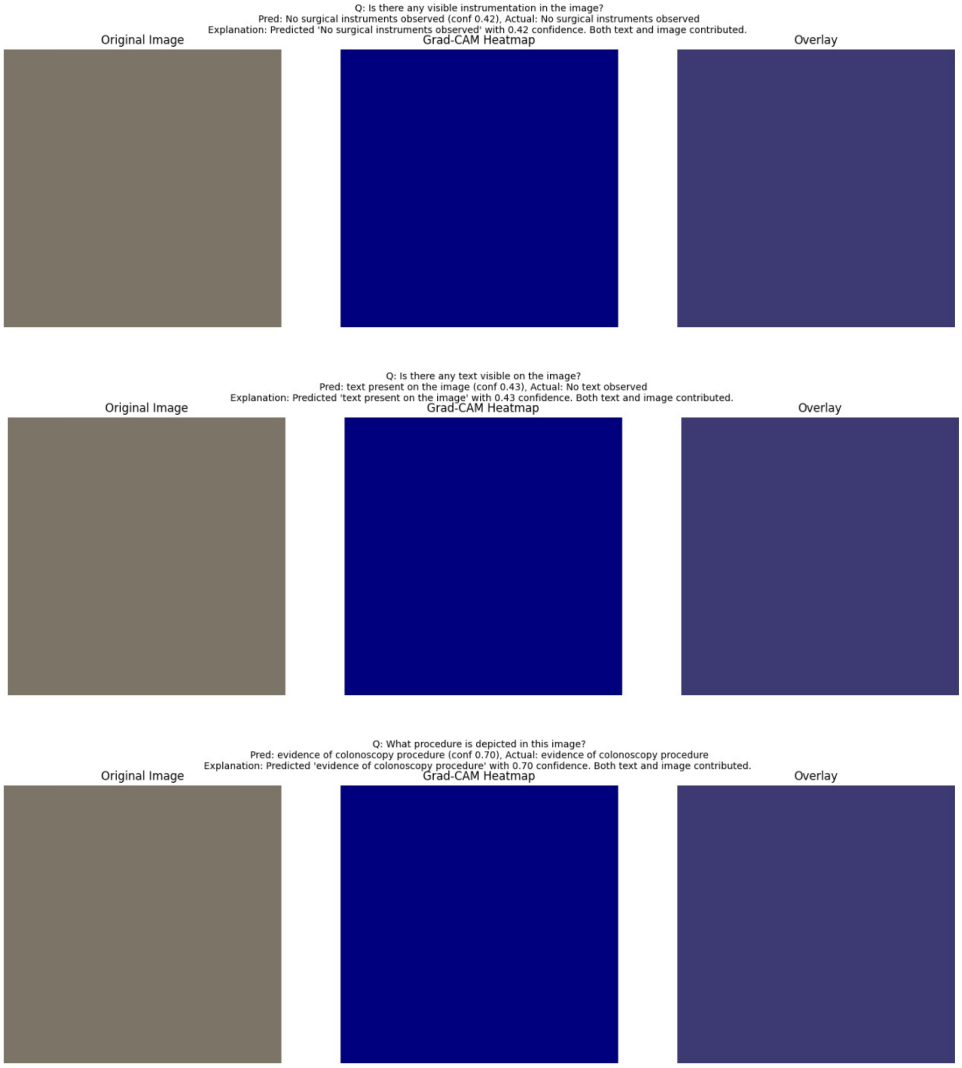
For interpretability, Grad-CAM [9] was applied to visualize attention maps, highlighting the image regions influencing predictions and ensuring clinical relevance.

4 Exploratory Data Analysis (EDA)

The **Kvasir-VQA-x1** dataset comprises six question categories — Yes/No, Single-Choice, Multiple-Choice, Color, Location, and Numerical Count — with Yes/No questions being the most frequent, resulting in class imbalance. Around 70% of answers contain fewer than three tokens, while reasoning and counting questions tend to be longer. All images were resized to 224×224 RGB format for CNN input, ensuring consistent visual quality. Correlation analysis revealed a strong relationship ($r \approx 0.95$) between question and answer lengths, but minimal correlation between image and text embeddings, confirming their complementary roles in multimodal learning. Techniques such as undersampling, answer normalization, and tokenization were applied to mitigate skewness in label distribution and standardize textual inputs.

Additionally, preliminary analysis included visualization of answer type distributions, question length histograms, and image quality checks. Data augmentation techniques like horizontal flips and slight rotations were applied to enhance model generalization. These steps ensured that both visual and textual modalities were properly preprocessed for effective training of the Ultra-Efficient VQA Model.

In summary, these observations highlight the diversity and imbalance present in the dataset, which directly influences model training and evaluation. The overall feature correlations are illustrated in the heatmap shown below.



5 Model Performance Comparison

Table 1 summarizes the performance of different model variations on the Kvasir-VQA-x1 dataset. We report official VQA metrics, including Precision, Accuracy, BLEU, METEOR, and ROUGE scores. Additional rows illustrate the impact of training ablations (e.g., scheduler, early stopping) and the integration of explainability.

Model Variant	**Dataset / Setup**	**Precision**	**Accuracy**	**BLEU**	**METEOR**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Baseline CNN+Text (No Scheduler)	Train/Test (Early Runs)	0.612	63.68%	0.642	0.812	0.854	0.776	0.853
CNN+Text + LR Scheduler	Train/Test	0.648	64.90%	0.670	0.840	0.865	0.800	0.865
CNN+Text + Early Stopping	Train/Test	0.650	65.00%	0.672	0.842	0.867	0.802	0.867
CNN+Text + LR Scheduler + Early Stop	Train/Test (Best Model)	0.661	66.07%	0.6865	0.8550	0.8769	0.8146	0.8769
CNN+Text + Explainability (Grad-CAM)	Post-hoc (Best Model)	–	66.07%	–	–	–	–	–

Table 1: Model Performance Comparison across different configurations, including training ablations and explainability evaluation.

From Table 1, we observe that combining both the learning rate scheduler and early stopping yields the best overall performance in terms of accuracy and text-based metrics (BLEU, METEOR, ROUGE). The integration of Grad-CAM provides visual explanations without impacting predictive performance, reinforcing the interpretability of the proposed approach.

6 Conclusion

We proposed the **Ultra-Efficient VQA Model**, an interpretable and computationally efficient baseline for Medico 2025 Subtask 1 using the Kvasir-VQA-x1 dataset. By combining a ResNet-18 CNN with a Bi-LSTM in an early-fusion framework, the model achieves competitive gastrointestinal VQA performance without heavy transformer architectures. Grad-CAM visualizations enhance interpretability, allowing clinicians to verify predictions. This work provides a reproducible foundation for trustworthy AI in GI imaging, and future research may explore hybrid models, lightweight multimodal transformers, and advanced explainability methods to further improve accuracy and transparency.

References

[1] Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). "Stacked Attention Networks for Image Question Answering." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[3] Tan, H., & Bansal, M. (2019). "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[4] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks." In *Advances in Neural Information Processing Systems (NeurIPS)*.

[5] Li, L., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2020). "VisualBERT: A Simple and Performant Baseline for Vision and Language." In *arXiv preprint arXiv:1908.03557*.

[6] Kim, W., Son, B., & Kim, I. (2021). "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision." In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[7] Abacha, A. B., et al. (2021). "Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering in the Medical Domain." In *CLEF 2021 Working Notes (CEUR Workshop Proceedings)*.

[8] He, X., Zhang, Y., & Wang, S. (2020). "PathVQA: 30000+ Questions for Medical Visual Question Answering." In *arXiv preprint arXiv:2003.10286*.

[9] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[10] Gautam, A., et al. (2025). "Medico 2025: Visual Question Answering for Gastrointestinal Imaging." *arXiv preprint arXiv:2508.10869*.