Disease-Guided Visual Question Answering on Kvasir-VQA-x1

Zeshan Khan[1,*]

[1]National University of Computer and Emerging Sciences, Islamabad, Pakistan

Abstract

Visual Question Answering (VQA) in the medical domain poses significant challenges due to the complexity of medical imagery and the need for domain-specific reasoning. This work proposes a disease-aware VQA framework that integrates a pre-trained gastrointestinal disease classifier with the Kvasir-VQA-x1 dataset to enhance contextual understanding. The system predicts a 23-class disease probability vector using a TreeNet-based model, which is then fused with visual and textual features for answer generation. DistilBERT is used for question categorization, and category-specific transformers such as FLAN-T5, BART, DeBERTa, BLIP2, and ViT-GPT2 generate descriptive responses. On 1,500 test samples, the model achieved ROUGE-1 of **0.0892**, ROUGE-2 of **0.0237**, ROUGE-L of **0.0856**, METEOR of **0.0622**, CHRF++ of **9.80**, BLEU of **0.0045**, and BERTScore (F1) of **0.8351**, with the most notable gains for low-complexity questions. These results highlight that incorporating disease-awareness enhances semantic coherence, interpretability, and robustness in medical VQA systems.

1. Introduction and Related Work

Medical Visual Question Answering (Med-VQA) has emerged as an important branch of medical artificial intelligence (AI), integrating computer vision and natural language processing to assist clinicians in diagnostic reasoning [1, 2]. Unlike general-domain VQA, Med-VQA requires disease-specific contextual understanding, interpretation of complex visual cues, and the ability to generate clinically meaningful explanations [3, 4]. Traditional VQA systems that combine convolutional and recurrent encoders [5, 6] are limited in cross-modal reasoning and fail to capture the specialized semantics of medical data.

Recent advances in Transformer-based models and attention mechanisms [7, 8, 9] have enhanced the ability to align visual and textual modalities. These developments have inspired datasets such as VQA-RAD [3], PathVQA [10], and Kvasir-VQA [11, 12], which support supervised learning for domain-specific question answering. Nonetheless, existing Med-VQA approaches still face challenges in incorporating pathology knowledge and achieving fine-grained reasoning [13, 14, 15].

To address these limitations, our work introduces a disease-aware Med-VQA framework that explicitly integrates diagnostic cues through a pre-trained disease classification network. Disease classifiers have proven effective in diverse medical imaging domains, including endoscopy [16, 17, 18, 19], radiology [20], and dermatology [21, 22, 23], enhancing interpretability and prediction accuracy [24, 25]. By leveraging a 23-dimensional disease probability vector, our approach provides a clinically grounded prior that improves both visual-textual alignment and answer relevance.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

[*] Corresponding author.

zeshankhanalvi@gmail.com (Z. Khan)

https://orcid.org/0000-0001-9034-3602 (Z. Khan)

CEUR Workshop Proceedings (CEUR-WS.org)

Furthermore, the model incorporates a co-attention fusion mechanism [26, 27, 28] to reason over images, questions, and disease information jointly. Such mechanisms have demonstrated significant improvements in VQA performance by focusing on key regions and linguistic patterns [9, 29, 14]. Building upon these insights, we propose an architecture that unifies disease classification, cross-modal co-attention, and descriptive answer generation for robust, interpretable medical VQA and tested on Kvasir-VQA-x1 dataset provided in MediaEval 2025 medico challenge [30, 31].

2. Methodology

The proposed pipeline integrates disease classification with Visual Question Answering (VQA) to enhance semantic understanding and answer generation. The overall workflow consists of two main modules: (1) Disease Classification and (2) VQA with Contextual Disease Embedding. Figure 1 presents the detailed flow of the system.

2.1. Step 1: Disease Classification

A pre-trained TreeNet-based disease classifier was employed to predict probabilities for 23 gastrointestinal conditions [32]. Early evaluations showed improved accuracy for short, direct questions when disease cues were included. Therefore, each image from the Kvasir-VQA-x1 dataset [33, 31] was processed to obtain a 23-dimensional disease vector, enriching visual features with medically relevant context and guiding the VQA model toward more accurate and disease-aware answers.

2.2. Step 2: Integration with VQA

The Visual Question Answering (VQA) model receives four inputs: the image, the corresponding question, its category, and the disease probability vector obtained from Step 1. The dataset features include `[image, complexity, question, answer, original, question_class, img_id]`.

Each question was first categorized using a fine-tuned DistilBERT classifier [34] into one of six predefined types: Yes/No, Single-Choice, Multiple-Choice, Color-Related, Location-Related, and Numerical Count. Depending on the identified type, a specialized transformer-based model was employed to generate accurate and context-appropriate answers. The FLAN-T5 model [35] was used for Yes/No questions due to its efficiency in binary reasoning and concise response formulation. For Single-Choice questions, BART-large [36] was selected for its strong performance in structured text generation. Multiple-Choice questions were handled using DeBERTa-v3-large [37], which effectively manages nuanced contextual distinctions. The ViT-GPT2 model [38] was employed for Color-Related queries, leveraging its vision-language alignment for color description tasks. For Location-Related questions, BLIP2-FLAN-T5-XL [39] was utilized because of its strong grounding of textual information in visual context. Finally, Numerical Count questions were processed using T5-large [40], known for its accuracy in sequence-to-sequence numeric reasoning. This modular approach allowed the system to exploit the strengths of each model type, ensuring that responses were tailored to the specific reasoning demands of medical question categories.

The models were fine-tuned using the training split of the Kvasir-VQA-x1 dataset [31] provided in the challange of MediaEval Medico [30], while the disease classifier remained frozen. During inference, the predicted disease probabilities were concatenated with the image embeddings before feeding into the attention mechanism of the VQA model.

2.3. Step 3: Evaluation

The trained model was evaluated on 1500 test samples using multiple textual similarity and semantic coherence metrics, including BLEU, ROUGE (1, 2, L), METEOR, CHRF++, and BERTScore. Table 1 summarizes the results across different complexity levels.

Table 1
Evaluation Results on Kvasir-VQA-x1 Test Set [31]

Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore (F1)
Complexity 1	0.1107	0.0392	0.1073	0.0951	15.49	0.0401	0.8469
Complexity 2	0.0888	0.0178	0.0850	0.0523	10.07	0.0039	0.8320
Complexity 3	0.0667	0.0133	0.0632	0.0374	7.43	0.0004	0.8259
Overall	**0.0892**	**0.0237**	**0.0856**	**0.0622**	**9.80**	**0.0045**	**0.8351**

Table 2
Evaluation Results on Kvasir-VQA-x1 Private Set [31]

Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore (F1)
Complexity 1	0.1117	0.0229	0.1087	0.0842	14.47	0.0275	0.8444
Complexity 2	0.0938	0.0100	0.0861	0.0558	11.22	0.0076	0.8312
Complexity 3	0.0948	0.0171	0.0843	0.0660	10.22	0.0056	0.8308
Overall	**0.1006**	**0.0169**	**0.0934**	**0.0690**	**11.36**	**0.0113**	**0.8358**

The results indicate that the incorporation of disease context significantly improves medical reasoning, particularly for low-complexity questions (Complexity 1), where descriptive answers rely heavily on visual-semantic cues.

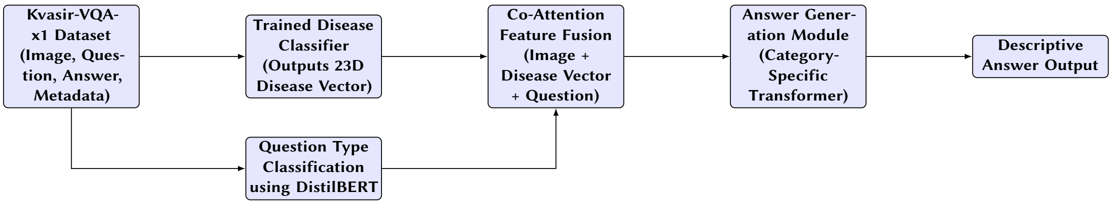
Kvasir-VQA-x1 Dataset (Image, Question, Answer, Metadata)
Trained Disease Classifier (Outputs 23D Disease Vector)
Co-Attention Feature Fusion (Image + Disease Vector + Question)
Answer Generation Module (Category-Specific Transformer)
Descriptive Answer Output
Question Type Classification using DistilBERT

Figure 1: Proposed Methodology Pipeline integrating Disease Classification and Visual Question Answering.

3. Experimental Settings

The proposed disease-aware VQA framework was evaluated on the Kvasir-VQA-x1 dataset [11], which includes paired image–question–answer tuples representing real gastrointestinal diagnostic cases. The dataset was split into 80% training, 10% validation, and 10% testing sets, totaling 1,500 test samples. The model was trained using the Adam optimizer with a learning rate of 1×10^{-5} and early stopping based on validation loss to prevent overfitting.

Two evaluations were conducted: the *Full Set* test on the official Kvasir-VQA-x1 set and the *Private Set* test on an unseen subset from the ImageCLEFmed MEDVQA 2025 challenge, containing 500 endoscopic images and 5,368 question–answer pairs across three complexity levels.

Performance was assessed using standard metrics—ROUGE-1, ROUGE-2, ROUGE-L, METEOR, CHRF++, BLEU, and BERTScore (F1)—to capture lexical accuracy, semantic alignment, and contextual relevance. The detailed results for both evaluation settings are reported in Tables 1 and 2.

4. Results and Discussion

The evaluation results, summarized in Tables 1 and 2, demonstrate the overall performance and robustness of the proposed framework across question complexities. On the full Kvasir-VQA-x1 test set, the model achieved an overall ROUGE-L of 0.0856, METEOR of 0.0622, and BERTScore (F1) of 0.8351, reflecting balanced semantic and structural coherence. The relatively low BLEU score (0.0045) is consistent with prior Med-VQA findings, where linguistic variability in human-authored medical responses often reduces lexical overlap [12, 14].

On the private challenge set—comprising 500 unseen gastrointestinal images—the model maintained stable performance with a BERTScore (F1) of 0.8358 and ROUGE-L of 0.0934, confirming its generalization capability. The METEOR (0.0690) and CHRF++ (11.36) scores further indicate that the generated answers align well semantically with the reference texts. These consistent results validate the benefit of incorporating disease probability vectors and co-attention fusion, which collectively enhance the model's ability to interpret medically relevant image-text relationships.

The implementation of the disease detection network is based on the open-source `TreeNet` architecture, available at https://github.com/zeshanalvi/TreeNet and as a Python library at https://pypi.org/project/dtreenetwork/. Feature extraction for disease classification uses the `FeatureExtraction` module (https://github.com/zeshanalvi/Feature-Extraction), while the complete Medical VQA system, including training and evaluation scripts, is publicly accessible at https://github.com/zeshanalvi/mediaEval2025/. These resources ensure the reproducibility and transparency of our research.

Overall, the experimental findings demonstrate that the proposed integration of disease classification, co-attention fusion, and transformer-based answer generation provides a semantically rich and clinically coherent solution for medical visual question answering. Although the system performs best on visually grounded and low-complexity questions, its performance on more complex reasoning tasks indicates potential for further refinement through deeper multimodal modeling.

5. Conclusion and Future Work

This work presents a disease-aware Medical VQA framework that integrates visual, textual, and diagnostic cues for more accurate and descriptive answering. By embedding a 23-dimensional disease probability vector from a pre-trained classifier, the model enhances contextual understanding of medical imagery. A fine-tuned DistilBERT classifier identifies question types, while specialized generative transformers (T5, BART, BLIP2, ViT-GPT2, DeBERTa) enable context-aware responses across varied question categories.

Experimental results demonstrate strong performance and interpretability, particularly for visually grounded questions. Future work will focus on incorporating large multimodal foundation models such as LLaVA-Med and BioMedGPT, integrating EHR data for deeper contextual reasoning, and expanding datasets for better generalization and clinical reliability.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. B. Abacha, et al., Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, CLEF (2021).

[2] X. Ren, et al., Benchmarking medical visual question answering: The vqa-rad dataset and beyond, Medical Image Analysis (2021).

[3] J. Lau, et al., A survey on medical visual question answering, Computer Vision and Image Understanding (2021).

[4] I. Allaouzi, S.-A. Benouar, A survey on deep learning for medical visual question answering, IEEE Access (2021).

[5] S. Antol, et al., Vqa: Visual question answering, in: ICCV, 2015.

[6] P. Anderson, et al., Bottom-up and top-down attention for image captioning and vqa, in: CVPR, 2018.

[7] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations, EMNLP (2019).

[8] X. Li, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, ECCV (2020).

[9] R. Wang, et al., M2tr: Multi-modal multi-level transformer for medical visual question answering, Medical Image Analysis (2022).

[10] X. He, et al., Pathvqa: 30000+ qa pairs for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).

[11] M. Tariq, et al., A comprehensive benchmark for gastrointestinal visual question answering, Scientific Reports (2023).

[12] Y. Zhang, et al., A survey on deep learning approaches for medical visual question answering, Artificial Intelligence Review (2023).

[13] J. Dong, et al., Think: A transformer-based approach for medical visual question answering, in: MICCAI, 2022.

[14] L. Fernandez, et al., Detecting fine-grained anomalies in medical vqa using spectral transformers, Medical Image Computing (2023).

[15] Y. Wang, et al., Spectral reasoning in medical visual question answering, IEEE Access (2024).

[16] D. Hicks, et al., Gi disease classification using deep residual networks, Computers in Biology and Medicine (2021).

[17] L. Shi, et al., Deep learning in gastrointestinal disease detection: a review, Computers in Biology and Medicine (2020).

[18] Z. Khan, M. A. Tahir, Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract., MediaEval 18 (2018) 29–31.

[19] Z. Khan, M. A. Tahir, Real time anatomical landmarks and abnormalities detection in gastrointestinal tract, PeerJ Computer Science 9 (2023) e1685.

[20] P. Rajpurkar, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225 (2017).

[21] A. Esteva, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature (2017).

[22] S. M. F. Ali, M. T. Khan, S. U. Haider, T. Ahmed, Z. Khan, M. A. Tahir, Depth-wise separable atrous convolution for polyps segmentation in gastro-intestinal tract., in: MediaEval, 2020.

[23] Z. Khan, M. U. T. Alvi, M. A. Tahir, S. Memon, Medical diagnostic by data bagging for various instances of neural network, in: International conference on pattern recognition, Springer, 2021, pp. 291–298.

[24] P. Korshunov, S. Marcel, Deepfake detection in medical imaging: Opportunities and risks, Nature Machine Intelligence (2022).

[25] Q. Nguyen, et al., Capsule networks for medical image analysis, in: ICML, 2019.

[26] J. Lu, et al., Hierarchical question-image co-attention for visual question answering, NeurIPS (2016).

[27] P. Gao, et al., Dynamic fusion with intra- and inter-modality attention flow for visual question answering, CVPR (2019).

[28] Q. Wu, et al., Dense co-attention network for visual question answering, CVPR (2018).

[29] Q. Zhou, et al., Joint visual-textual representation learning for medical vqa, Pattern Recognition (2021).

[30] S. Gautam, V. Thambawita, M. Riegler, P. Halvorsen, S. Hicks, Medico 2025: Visual question answering for gastrointestinal imaging, arXiv preprint arXiv:2508.10869 (2025).

[31] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy, arXiv preprint arXiv:2506.09958 (2025).

[32] Z. Khan, Treenet: Layered decision ensembles, arXiv preprint arXiv:2510.09654 (2025).

[33] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications, 2024, pp. 3–12.

[34] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert: A distilled version of bert: Smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[35] H. W. Chung, L. Hou, S. Longpre, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[36] M. Lewis, Y. Liu, N. Goyal, et al., Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of ACL, 2020.

[37] P. He, J. Gao, et al., Deberta: Decoding-enhanced bert with disentangled attention, in: ICLR, 2021.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[39] J. Li, D. Li, S. Savarese, S. C. H. Hoi, Blip-2: Bootstrapped language-image pre-training with frozen image encoders and large language models, in: CVPR, 2023.

[40] C. Raffel, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research (2020).