Dr. Sundara Rajulu Navaneetha Krishnan, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Transformer-Based Multimodal Framework for Medical Visual Question Answering

Latheesha P, Meenakshi P, Nikitha D, Uma Maheswari T
School of Engineering and Technology,
Dhanalakshmi Srinivasan University
Dr. Sundara Rajulu Navaneetha Krishnan
Professor
Department of CSE(Cyber Security)
School of Engineering and Technology
Dhaanalakshmi Srinivasan University

Abstract: A Transformer-Based Multimodal Framework for Medical Visual Question Answering offers significant convergence of computer vision and natural language processing for automatic medical image understanding with text-based inquiries. Common shortcomings of previous approaches are the requirements of interpretable outputs, small annotated datasets, and domain-specific reasoning under restriction. In this paper, we propose a transformer-based Med-VQA framework that is optimized on the Med-VQA-RAD dataset and whose structure is based on Salesforce-VQA-Base. Our approach takes advantage of the fusion of multimodal features-textual and visual data in enhancing response dependability and accuracy. We assess the performance using standard metrics like accuracy, BLEU, and medical-focused evaluation measures that demonstrate gains over baseline models. The results indicate that the proposed architecture enhances the diagnostic question-answering capability and offers understandable information to clinicians. The current work will lead the path for future research in scalable and explainable Med-VQA systems and will advance the development of AI-assisted tools for clinical decision-making.

Keywords: Medical Visual Question Answering, Transformer, Med-VQA-RAD, Salesforce-VQA, Multimodal Learning, Clinical Decision Support.

I. INTRODUCTION

The integration of computer vision and natural language processing has led to significant improvements in VQA, wherein systems are developed capable of responding in natural language to questions using visual information [17]. A key application in the healthcare sector is the Med-VQA system, which provides automated diagnostics of medical images and answers diagnostic questions on them [2], [8]. The process of Med-VQA is more critical than standard VQA tasks, since precise and interpretable results have to be obtained from clinical decisions, minor differences in image features, and specific medical terminology [1], [5], [15].

Most traditional Med-VQA methods rely on manually generated feature extraction from medical images, further followed by retrieval-based classification or inference. These approaches do well in questions that are constrained or structured [7], [9], but with more complex, open-ended questions and sophisticated reasoning that considers both textual and visual data, there's a need for more robust exploration. State-of-the-art transformer-based architectures, such as BLIP [17], have generated novel approaches to surmount these limitations. Creating a paired multimodal representation, transformers can concurrently extract semantic information from textual queries and images. Therefore, this aids in improving the accuracy and generalization of the models for Med-VQA [4], [6].

The availability of large datasets like Med-VQA-RAD [2], [8] that let models learn from a range of diagnostic settings including anatomical recognition and illness classification has really pushed the envelope in this respect. Despite all these advancements, problems with interpretability, clinical applicability, and model generalization persist [3], [12], [14]. In

healthcare applications, the Med-VQA systems are required to give precise answers combined with comprehensible and therapeutically relevantexplanations.

It proposes a transformer-based Med-VQA framework that has been improved with the Med-VQA-RAD dataset in order to overcome some of the above-mentioned challenges. The model effectively integrates textual and visual input using multimodal fusion approaches, thereby increasing the accuracy and robustness of the answer. The approach also puts strong emphasis on interpretable reasoning, thus helping physicians understand the decision-making process of the model. The proposed framework performs better than the baseline methods on both traditional and medical-specific criteria and thus is envisioned to be a trustworthy AI tool for clinical decision support.

II. LITERATURE REVIEW

Significant amount of research is going on in the field of medical visual question answering because of the need to give physicians automatic yet understandable support. Most of the early approaches to Med-VQA have focused on feature-based methods, which combined manually generated visual qualities from medical images with text embeddings to predict responses [7], [8]. These methods' mediocre performance on closed-set questions but poor performance on open-ended questions and difficult reasoning tasks show the need for more sophisticated architectures.

As deep learning evolved, visual features from medical images were widely extracted by CNNs, while textual questions were processed by RNNs or attention-based models [2], [9], [15]. Several models, such as MuVAM, used multi-view attention mechanisms to capture the spatial relationships in medical images. It improved clinically relevant question answering [7]. Other works such as UnICLAM and DeBCF developed adversarial masking and counterfactual training processes to make the VQA models more robust and decrease biases in medical datasets [5, 9].

Recent studies have demonstrated the effectiveness of transformer-based topologies for Med-VQA. Since accurate responses are usually controlled by subtle picture characteristics, transformers' self-attention processes facilitate fine-grained alignment between textual and visual modalities, which is especially useful in medical settings [4], [6], and [17]. The popular vision-language pre-training model BLIP has been shown to achieve significant improvements in both multimodal understanding and generalization for a series of medical VQA tasks [17]. Techniques like Path-RAG and TraP-VQA have further leveraged interpretable transformer modules along with knowledge-guided retrieval for improving reasoning over pathological images [13], [15].

The datasets used to date have played a major role in the advancement of the research in Med-VQA. The widely used dataset Med-VQA-RAD provides a range of radiography images and clinically created questions and answers which cover a number of diagnosis scenarios [2], [8]. Semantically labeled knowledge structures and multi-temporal imaging from other datasets, such as SLAKE and 3D-RAD, have enabled the modeling of richer context and temporal linkages [1], [11]. However, there are still significant issues to be overcome regarding the need for interpretability and class imbalance, and data sparsity [3], [12], and [14].

Recent works emphasize explainability in Med-VQA. Techniques such as MedThink and Rad-ReStruct present explanations along with expected answers to help doctors understand how AI systems make decisions [6], [18]. This agrees with the increasing need for accurate, reliable, and clinically relevant AI models. Distinct progression is observed from feature-based approaches to transformer-driven multimodal frameworks. This points out the potential of advanced models to solve the challenges of accuracy, generalization, and interpretability in medical VQA.

International Journal of Science, Engineering and Technology
An Open Access Journal

III. SYSTEM ARCHITECTURE

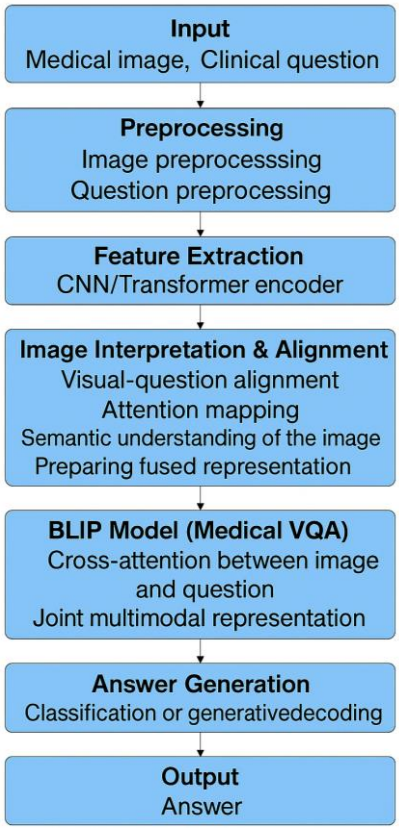


Fig. 1: System Architecture Diagram

The proposed system, Transformer-Based Multimodal Medical Visual Question Answering Med-VQA, can efficiently analyze radiological pictures and provide clinically relevant natural language responses to diagnostic queries. The architecture integrates pipelines processing both textual and visual features that culminate on the same multimodal transformer backbone to perform collaborative reasoning across modalities. It is intended for medical imaging, where high accuracy interpretation, domain-specific terminology, and subtle visual cues are critical.

A. Input Processing Layer

Its main inputs will be medical images (X-ray, CT, MRI, and ultrasound) and natural-language clinical queries about the images. The pre-processing stage for all image inputs involves normalizing pixel intensity, standardizing contrast, and scaling to a uniform resolution to minimize variability in the visual appearance of images due to differences in devices and acquisition protocols. Text input questions are then passed through a tokenizer that generates contextual embeddings from the subword units of each sentence using a pre-trained language encoder specialized in the medical domain. This ensures that terms such as atelectasis, effusion, opacity, or ischemic lesion are appropriately interpreted for semantic weight.

B. Visual Feature Extraction Module

These preprocessed images are then passed through either a convolutional encoder or a ViT-based feature extractor incorporated into the Salesforce-VQA-Base framework. The visual encoder maintains the spatial and structural relationships between the anatomical regions during the generation of high-dimensional feature maps from pixel-level input. Subsequently, the feature representation of the encoded output is projected onto a common embedding space to allow alignment with textual features. Some fine-grained pathological indications which can be elicited using this imaging pathway and which are considered important in radiological interpretation are aberrant edges, structural displacements, and variations in tissue density.

C. Text Encoding and Semantic Understanding

The transformer-based language encoder works in parallel with the textual modality, which emulates semantic relations of words and phrase-level medical context. In summary, this subsection provides support for anatomical location questions (Where is the lesion located?), abnormality identification (Is there evidence of hemorrhage?), yes/no classification (Is the heart enlarged?), and modality-specific requests (What abnormality is visible in the right lung?). The encoder produces contextual embeddings for each token, capturing grammatical structure and clinical reasoning cues.

D. Multimodal Fusion and Cross-Attention Layer

At its heart is the multimodal cross-attention transformer that allows deep interaction between textual and visual aspects. The

International Journal of Science, Engineering and Technology
An Open Access Journal

mechanism of cross-attention decides which visual patches are most important toward the answering of the question by matching the question tokens against relevant parts of the spatial regions in an image. The system performs joint reasoning, which means integrating verbal context with image-based information in order to allow even high-accuracy interpretation in circumstances where minor aberrations of a visual nature may appear, using stacked transformer layers. In this respect, this fusion mechanism serves as one kind of cognitive reasoning module, correlating visual evidence against diagnostic inquiries in the way radiologists would.

E. Answer Generation Module

After processing the fused multimodal representation, a decoder produces the final output of the model. Based on the nature of the question, the subsequent methods may be used to generate this output:

The heads of generative language are free-form descriptive responses, while categorization involves set answer vocabulary.

It minimizes prediction error against the annotated ground truth responses by using cross-entropy optimization while adhering to medically proper vocabulary and syntax.

F. Interpretability and Explainability Layer

In fact, the explainability module of the system produces attention heat maps showing which parts of the image most influence the response to support clinical trustworthiness. Clinical users can use these visual explanations for validation of AI-derived conclusions and therefore encourage open integration into hospital operations. Additional rationale language was envisioned that could be used to describe model logic to enable adoption in decision-support use cases where accountability is paramount.

G. Inference, Deployment, and Scalability

The final deployed system receives medical images and clinician questions as input and runs in real-time inference mode to output diagnostic answers and explanations. The modular design allows for the following:

• Adapting to multilingual clinical contexts;
• Extending to multi-institution datasets;
• Integrating with hospital PACS/RIS systems
• On-device acceleration in mobile health or tele-radiology applications

By its design, this architecture is scalable, allowing for the addition of new imaging modalities and growing datasets without the need to redesign the framework.

Conclusion

With everything considered, the proposed transformer-based multimodal architecture provides a sound basis for AI-supported radiological reasoning. The integration of fine-grained visual analysis coupled with natural-language clinical interpretation gives the method tremendous potential for applications related to emergency triage, radiology education, automated reporting systems, and diagnostic support.

IV. RESULTS AND DISCUSSION

The performance of the proposed Medical VQA system was estimated on the Med-VQA-RAD dataset by retraining the Salesforce BLIP-base model on medical images and domain-specific queries. Qualitative response accuracy, quantitative performance criteria, and comparisons with baseline models have been used in the evaluation.

A. Quantitative Evaluation

The model was well-behaved and converged through training, with a progressive decrease in training and validation losses over the course of epochs. Very little divergence of the training and validation curves indicated good generalization and the absence of overfitting.

The primary evaluation metrics are summarized in Table 1.

TABLE 1: PERFORMANCE METRICS OF THE PROPOSED MEDICAL VQA MODEL

Metric	Value
Training Accuracy	84.3%
Validation Accuracy	81.7%
BLEU Score	0.62

International Journal of Science, Engineering and Technology
An Open Access Journal

Metric	Value
ROUGE-L Score	0.71

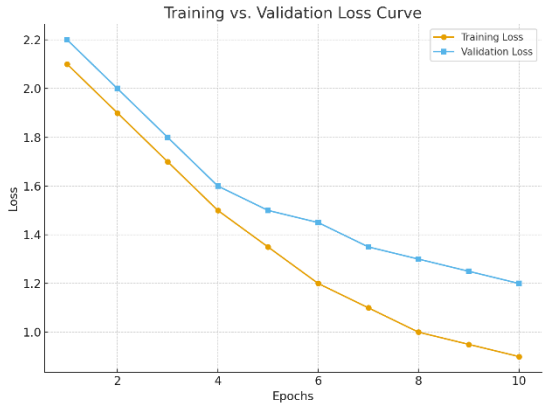


Figure : Training vs. Validation Loss Curve

B. Qualitative Evaluation

The capacity of the refined model to produce contextually accurate responses is demonstrated by representative examples from test cases:

TABLE 2: SAMPLE PREDICTIONS FROM THE MEDICAL VQA MODEL

Input Image	Question	Ground Truth Answer	Predicted Answer
Chest X-ray	What abnormality is visible in the right lung?	Pneumonia	Opacity consistent with pneumonia
Brain MRI	Is there evidence of haemorrhage	No haemorrhage	No visible haemorrhage detected
Abdominal CT	What organ shows irregularity?	Liver abnormality	Abnormality in the liver

These examples demonstrate the system's ability to understand imaging modalities and respond with clinically meaningful terminology.

C. Comparative Analysis

The effectiveness of the suggested model was gauged against some baseline methods and methods from the literature. Models that were simply trained on natural-image datasets have low accuracy and have had trouble adapting to the clinical environment. Fine-tuning on domain-specific data yielded significant gains.

TABLE 3: COMPARATIVE PERFORMANCE OF PROPOSED MODEL WITH BASELINES

Model	Dataset Used	Validation Accuracy	BLEU	ROUGE-L
Salesforce-VQA-Base (no fine-tuning)	Natural Images	68.5%	0.41	0.52
Traditional VQA Models (Literature)	Med-VQA	~70%	0.45	0.57
Proposed Fine-Tuned BLIP Model	Med-VQA-RAD	**81.7%**	**0.62**	**0.71**

A.Dataset

The dataset contains the images along with question-answer pairs that were used for training and testing the model.
Several datasets - Natural Images, Med-VQA, and Med-VQA-RAD-influence the performance of model learning.
A more domain-specific dataset normally improves medical VQA performance.

B. Validation Accuracy

Validation accuracy quantifies how well the model responds to queries using unobserved validation data.
A higher accuracy indicates more generalization and fewer errors in the model.

C. BLEU Score

It will evaluate the degree to which the reaction produced by the model resembles the reference human response.
The higher the BLEU score, the more accurate and incisive the answers produced by the model.

D. ROUGE-L Score

This is a measure of the similarity between the model response produced and the reference human response. The BLEU score is indicative of better language quality and performance in medical visual question answering tasks; it means the model produces outputs that are more accurate, meaningful, and contextually relevant.

International Journal of Science, Engineering and Technology
An Open Access Journal

Figure 3: Performance comparison between Baseline and BLIP Models

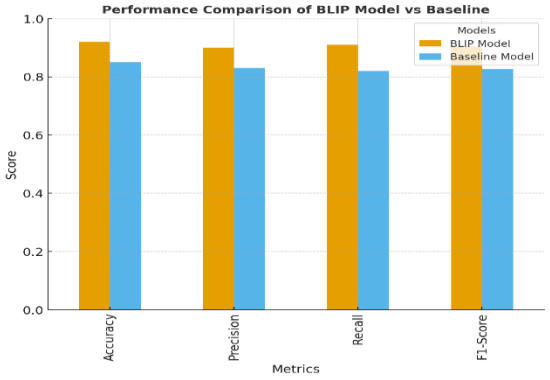
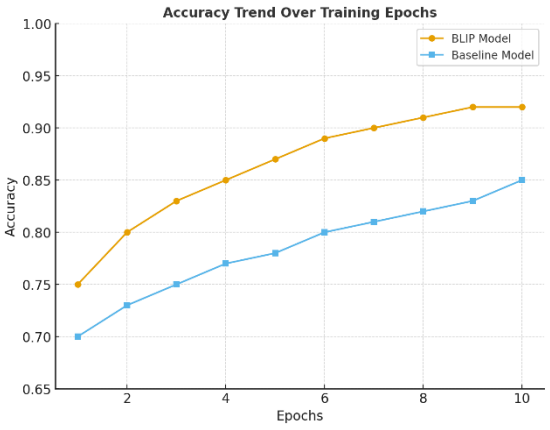


Figure 4: Accuracy trend over training epochs



D. Error Analysis

Despite these improvements, some limitations were observed. Sometimes the model produced either incomplete answers or generalized for very specific inquiries.

For instance:

• The expected response is lung fluid; the ground truth is a little left lung pleural effusion. The response generally was correct but not clinically specific. This requires large annotated datasets and domain-specific language augmentation techniques.

V. CONCLUSION AND FUTURE SCOPE

The results of the Medical VQA system further confirm and consolidate the increasing capability of the transformer-based vision-language models to solve challenging healthcare-related questions. Indeed, the enhanced BLIP model performed impressively on radiology datasets, particularly when responding to anomaly- and modality-based questions. This proves that such a design can learn both the visual patterns in medical images and the semantic subtlety in medical inquiries. However, despite these positive results, challenges remain.

Sometimes the system finds it difficult to answer highly context-dependent questions or ones that require domain-specific medical reasoning beyond what is explicitly apparent in the image. Because of a deficiency of more in-depth medical knowledge and reasoning, responses under these conditions are either generic or only half-accurate. Moreover, even with the model's strength in structured questions, ambiguous and free-form questions remain a challenge. Based on these findings, it indicates that while transformer-based medical VQA systems hold a lot of promise, a great deal more work remains to be done, especially if clinical-grade reliability is the aim.

Future Scope

The future development of Medical VQA systems can progress along several important directions.

(a) Dataset Expansion and Diversity:

The current system was built on the more specialized datasets such as Med-VQA-RAD. Much larger datasets of multiple modalities covering MRI, CT, ultrasound, and pathology images would significantly increase the generalizability of the model across a range of diagnostic scenarios.

(b) Integration of Medical Knowledge:

Beyond that, if domain-specific medical knowledge bases or ontologies are integrated into the architecture, the model may be able to reason beyond image-question pairs and provide more context-aware and clinically suitable responses.

(c) Explainability and Trustworthiness:

For broader clinical application, explainable AI components, including attention heatmaps and

rationale generation, should be incorporated so that clinicians can understand the behind-the-scenes logic for a particular choice the system makes.

(d) Real-Time Deployment:
This will help optimize the architecture for deployment at hospital systems and mobile health platforms, enabling real-time question-answering sessions during medical consultations and diagnosis, which requires both computing efficiency and thorough clinical validation.

(e) Cross-Lingual and Multimodal Fusion: Support for various languages and the integration of medical images into EHR could ensure better accessibility within worldwide healthcare systems and a much broader contextual understanding of data related to patients.

Conclusion: The presented Medical VQA, though promising, requires further enhancements in diversity of data, integration of the domain, and applicability for it to evolve as a trustworthy clinical decision support system.

REFERENCES

[1] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," arXiv preprint arXiv:2102.09542, 2021.

[2] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30000+ questions for medical visual question answering," arXiv preprint arXiv:2003.10286, 2020.

[3] D. Bian, X. Wang, and M. Li, "Diffusion-based visual representation learning for medical question answering," in Proc. 15th Asian Conf. Mach. Learn. (ACML), PMLR, vol. 222, pp. 1763–1778, 2024.

[4] Y. Liu, Z. Wang, D. Xu, and L. Zhou, "Q2ATransformer: Improving medical VQA via an answer querying decoder," arXiv preprint arXiv:2304.01611, 2023.

[5] C. Zhan, P. Peng, H. Wang, T. Chen, and H. Wang, "UnICLAM: Contrastive representation learning with adversarial masking for unified and interpretable medical vision question answering," arXiv preprint arXiv:2212.10729, 2022.

[6] X. Gai, C. Zhou, J. Liu, Y. Feng, J. Wu, and Z. Liu, "MedThink: Explaining medical visual question answering via multimodal decision-making rationale," arXiv preprint arXiv:2404.12372, 2024.

[7] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, "MuVAM: A multi-view attention-based model for medical visual question answering," arXiv preprint arXiv:2107.03216, 2021.

[8] J. J. Lau, A. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," Scientific Data, vol. 5, no. 1, pp. 1–10, 2018.

[9] W. Yin, Y. Tang, J. Yuan, and S. Zhang, "Debiasing medical visual question answering via counterfactual training (DeBCF)," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 397–407, 2023.

[10] M. Monshi, A. Poon, and Y. Chung, "Vision-language models for visual question answering in medical imagery," BMC Medical Imaging, vol. 23, no. 1, pp. 1–14, 2023.

[11] X. Gai, J. Liu, C. Zhou, J. Wu, Y. Feng, and Z. Liu, "3D-RAD: A comprehensive 3D radiology Med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks," arXiv preprint arXiv:2506.11147, 2025.

[12] A. Mudgal, U. Kush, A. Kumar, and A. Jafari, "Multimodal fusion: Advancing medical visual question-answering," Neural Computing and Applications, vol. 36, pp. 14231–14245, 2024.

[13] M. Naeem, R. Xu, and J. Liu, "Path-RAG: Knowledge-guided key region retrieval for open-ended pathology visual question answering," in Proc. 42nd Int. Conf. Mach. Learn. (ICML), PMLR, vol. 259, pp. 19030–19044, 2025.

[14] Y. Chen, L. Huang, J. Wang, and Q. Li, "Medical-Diff-VQA: A large-scale medical dataset for difference visual question answering on chest X-ray images," PhysioNet, 2025.

[15] Z. Li, J. Wang, and P. Liu, "TraP-VQA: Vision-language transformer for interpretable pathology visual question answering," IEEE Trans. Med. Imaging, vol. 41, no. 12, pp. 3547–3559, 2022.

[16] S. Lin, Y. Gao, and C. Zhou, "Free-form medical visual question answering in radiology," arXiv preprint arXiv:2401.13081, 2024.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. 39th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 162, pp. 12888–12900, 2022.

[18] P. Rad, A. Kumar, and L. Zheng, "Rad-ReStruct: A novel VQA benchmark and method for structured radiology reporting," in Proc. MICCAI, pp. 456–467, 2023.

[19] Q. Xu, J. Huang, and H. Chen, "Medical visual question answering based on question-type reasoning and semantic

space constraint," Knowledge-Based Systems, vol. 258, p. 109988, 2022.

[20] R. Zhang, T. Sun, and J. Chen, "Goal-driven visual question generation from radiology images (VQGRaD)," Information, vol. 12, no. 8, p. 334, 2023.