Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

SwinVQA: A Transformer Framework for Medical Visual Question Answering

Keyur Vaitha[1*], Jay Korat[2], Anand Akbari[3], Chinmay Raut[4]

[1*]Department of Computer Science and Engineering (Data Science), Dwarkadas Jivanlal Sanghvi College of Engineering, keyurvaitha21@gmail.com
[2]Department of Computer Engineering, Dwarkadas Jivanlal Sanghvi College of Engineering, koratjay22@gmail.com
[3]Department of Computer Engineering, Dwarkadas Jivanlal Sanghvi College of Engineering, anandakbari9@gmail.com
[4]Asst. Professor, Dept. of Computer Engineering, Dwarkadas Jivanlal Sanghvi College of Engineering, chinmay.raut@djsce.ac.in

Article History:

***Received**: 23-04-2025*

***Revised:** 29-05-2025*

***Accepted:** 08-06-2025*

Abstract:
The integration of advanced computer vision and natural language processing techniques in medical image analysis presents significant challenges due to the complexity and high stakes of diagnostic interpretation. This paper introduces SwinVQA, a novel medical visual question answering framework that leverages the hierarchical architecture of Swin Transformers to address limitations in existing systems. By employing shifted window partitioning and patch merging techniques, SwinVQA efficiently processes high-resolution medical images while simultaneously capturing both local details and global context essential for accurate diagnosis. We implement a sophisticated cross-modal attention mechanism that effectively aligns visual features with clinical queries, enhancing the model's reasoning capabilities. The framework is evaluated on established medical VQA datasets, including SLAKE and VQA-RAD, demonstrating improved performance across various question types and imaging modalities. Additionally, we introduce beam search optimization for answer generation, resulting in more contextually appropriate and diagnostically accurate responses. Experimental results show that SwinVQA significantly outperforms baseline models in both computational efficiency and diagnostic accuracy. This research advances the field of AI-assisted medical image analysis by providing a more robust and clinically relevant solution that bridges the gap between technological capabilities and practical healthcare applications.

***Index Terms*—**Medical Visual Question Answering, Swin Transformer, Hierarchical Feature Learning, Cross-Modal Fusion, Beam Search, Healthcare AI

I. INTRODUCTION

As the complexity and volume of medical imaging continue to increase, healthcare professionals face growing challenges in efficiently interpreting diagnostic data. Visual Question Answering (VQA) in the medical domain—a task aimed at automatically generating accurate textual responses to questions about medical images—emerges as a promising solution to enhance diagnostic capabilities and streamline clinical workflows. Despite advances in general VQA systems, medical applications present unique challenges that demand specialized approaches.

Medical Visual Question Answering (Med-VQA) systems must process high-resolution medical images with fine-grained details critical for diagnosis, understand complex anatomical structures across multiple scales, and generate responses with the precision required in clinical settings. Existing approaches, particularly those based on standard Vision Transformers (ViTs), face significant limitations including computational inefficiency when processing large medical images, inability to simultaneously capture local abnormalities and global anatomical context, and poor integration of visual and linguistic information.

The computational burden of traditional transformers, which scales quadratically with image size, renders them impractical for high-resolution medical images. Furthermore, the fixed partitioning scheme employed by standard ViTs limits their ability to effectively model hierarchical features essential for understanding both fine-grained details (e.g., small lesions) and broader anatomical structures. These limitations, coupled with inadequate fusion mechanisms for aligning visual and

textual modalities, result in systems that fall short of the accuracy and reliability required for clinical applications.

To address these challenges, we propose SwinVQA, a novel framework that leverages the hierarchical architecture of Swin Transformers for medical visual question answering. The Swin Transformer architecture offers key advantages through its shifted window partitioning mechanism and hierarchical design, enabling efficient processing of high-resolution images while capturing multi-scale features. By adapting this architecture specifically for medical imaging analysis and integrating it with advanced cross-modal fusion techniques, our system overcomes the limitations of existing approaches. The main contributions of this paper are:

- A hierarchical Swin Transformer-based vision encoder that efficiently processes high-resolution medical images through shifted window partitioning and patch merging, enabling multi-scale feature extraction critical for medical diagnosis.
- A sophisticated cross-modal fusion framework that effectively aligns visual features with textual queries, enhancing the model's ability to provide contextually relevant and diagnostically accurate answers.
- Implementation of beam search optimization for answer generation, significantly improving the quality and clinical relevance of generated responses compared to greedy decoding methods.
- Comprehensive evaluation on established medical VQA datasets, demonstrating superior performance in both computational efficiency and diagnostic accuracy compared to baseline approaches.

Our experimental results validate the effectiveness of SwinVQA in addressing the unique challenges of medical visual question answering, offering a promising solution for enhancing diagnosis and clinical decision-making in healthcare settings.

II. RELATED WORK

A. Visual Question Answering

Visual Question Answering (VQA) systems have evolved significantly since the seminal work of Antol et al. [1], who introduced the foundational VQA dataset and established baseline approaches for integrating visual and textual information. Early systems typically employed a two-channel approach: one channel processing images through convolutional neural networks (CNNs) like VGGNet, and another channel handling questions using either bag-of-words models or recurrent architectures such as Long Short-Term Memory (LSTM) networks. Xu and Saenko [2] advanced the field by introducing question-guided spatial attention mechanisms to highlight relevant image regions when answering queries. Their approach demonstrated improved performance on spatial reasoning tasks, though it lacked the ability to effectively integrate external knowledge sources. This limitation remained a persistent challenge in subsequent VQA systems, particularly when addressing complex queries requiring domain-specific knowledge.

The introduction of transformer architectures marked a significant advancement in VQA systems. Dosovitskiy et al. [3] proposed Vision Transformers (ViTs), which divided images into fixed-size patches processed through transformer encoders, demonstrating competitive performance on image recognition tasks. However, these models required substantial training data and computing resources to match or exceed CNN performance, limiting their practical application in specialized domains.

B. Medical VQA

The application of VQA techniques to medical imaging presents unique challenges beyond those of general VQA systems. He et al. [7] introduced PathVQA, one of the first specialized datasets for medical VQA, containing pathology images paired with clinically relevant questions. Their work highlighted the importance of domain-specific pretraining for achieving acceptable performance in medical applications.

Liu et al. [5] developed SLAKE, a semantically-labeled knowledge-enhanced dataset specifically designed for medical VQA. This dataset addresses the critical need for structured annotations and

knowledge integration in medical image analysis, providing a valuable benchmark for evaluating Med-VQA systems. Their subsequent work [6] proposed a conditional reasoning framework that improves performance through contrastive learning, achieving better alignment between visual features and clinical queries.

Abacha et al. [8] conducted comprehensive evaluations of various Med-VQA approaches through the VQA-Med task at ImageCLEF, identifying key challenges in medical image interpretation and response generation. Their findings emphasize the need for models that can accurately interpret varied imaging modalities and produce clinically precise answers, particularly for diagnostic questions.

C. Transformer Architectures for Image Analysis

Recent advances in transformer architectures have addressed many limitations of standard ViTs. Zhang et al. [4] integrated transformer components with YOLO detection frameworks, demonstrating improvements in object detection tasks. However, their approach still faced computational challenges when applied to high-resolution images.

Liu et al. [9] introduced the Swin Transformer, a hierarchical vision transformer that computes self-attention within local windows while allowing cross-window connections through shifted window partitioning. This architecture significantly reduces computational complexity from quadratic to linear with respect to image size, making it more suitable for highresolution image processing.

Hu et al. [10] further enhanced the Swin Transformer through their V2 variant, addressing training instability, resolution gaps between pretraining and fine-tuning, and data efficiency concerns. Their innovations included a residualpost-norm method with cosine attention and a log-spaced continuous position bias approach, enabling more effective transfer of models pretrained on low-resolution images to tasks involving high-resolution inputs.

While these advancements have shown promise in general computer vision tasks, their application to medical VQA remains relatively unexplored. Our work bridges this gap by adapting the hierarchical architecture of Swin Transformers specifically for medical image analysis and question answering, addressing the unique challenges of this domain.

III. PROBLEM DEFINITION AND METHODOLOGY

A. Problem Statement

The medical visual question answering task involves generating accurate and clinically relevant text responses to questions about medical images. Formally, given an input medical image I and a question Q, the goal is to produce an answer A that correctly addresses the query based on the visual content. This task is particularly challenging in the medical domain due to:

- The high resolution and complexity of medical images, which contain fine-grained details critical for diagnosis
- The need to simultaneously capture local abnormalities and global anatomical context
- The requirement for domain-specific knowledge integration and contextual understanding
- The high stakes of medical diagnosis, demanding greater precision and reliability than general VQA applications

Current approaches face significant limitations in addressing these challenges, particularly in terms of computational efficiency, multi-scale feature representation, and effective crossmodal fusion.

B. SwinVQA Architecture

Our proposed SwinVQA framework consists of three main components:

1) A hierarchical image encoder based on the Swin Transformer
2) A cross-modal fusion module for integrating visual and textual features
3) An answer generation module with beam search optimization

1) Hierarchical Image Encoder: At the core of our framework is a hierarchical image encoder that leverages the Swin Transformer architecture. Unlike standard Vision Transformers that process images as a sequence of fixed-size patches with global self-attention, the Swin Transformer introduces two key innovations: shifted window partitioning and patch merging.

Window Partitioning for Computational Efficiency: To address the quadratic complexity of global self-attention, we employ window partitioning, dividing the image into nonoverlapping windows. This approach reduces computational complexity from $O((H \times W)^2)$ to $O(M^2 \times H \times W)$, where $H \times W$ is the number of patches and $M \times M$ is the window size:

$x \in \mathrm{R}B \times H \times W \times C$ reshape/view$\longrightarrow x' \in \mathbb{R}^{B \times \frac{H}{M} \times M \times \frac{W}{M} \times M \times C}$ (1)_ _

This reshaped tensor is then permuted to facilitate windowbased self-attention:

$x'' \in \mathrm{R}(B \times \frac{H}{M} \times \frac{W}{M}) \times M \times M \times C$ (2)

Shifted Window Partitioning: To enable cross-window connections and capture global context, we implement a shifted window partitioning scheme that alternates between regular and shifted window configurations across transformer layers:

$x' = \mathrm{roll}(x, \mathrm{shifts} = (-S, -S), \mathrm{dims} = (1,2))$ (3) where S represents the shift size. After applying selfattention within these shifted windows, the feature map is restored to its original alignment:

$x_{\mathrm{restored}} = \mathrm{roll}(x_{\mathrm{shifted}}, \mathrm{shifts} = (S,S), \mathrm{dims} = (1,2))$ (4)

Patch Merging for Hierarchical Representation: To capture multi-scale features, we implement patch merging operations that progressively reduce spatial resolution while increasing feature dimensionality:

$x \in \mathrm{R}B \times H \times W \times C$ reshape/view$\longrightarrow x' \in \mathbb{R}^{B \times \frac{H}{2} \times \frac{W}{2} \times 4C}$ _ _ norm + reduction (5)

This hierarchical structure enables the model to learn representations at multiple scales, capturing both fine-grained details and global context essential for medical image interpretation.

2) Cross-Modal Fusion: To effectively integrate visual and textual information, we implement a sophisticated cross-modal fusion module. This module aligns features from the hierarchical image encoder with textual representations of the question. Textual Encoding:

The question Q is tokenized and processed through a BERTbased encoder to obtain contextual text embeddings:

$C = \{c_1, c_2, ..., c_N\}, c_j \in \mathrm{R}^d$ (6)

where N is the maximum sequence length and d is the embedding dimension.

Multimodal Alignment:

Visual and textual features are combined through concatenation and enhanced with modality-specific segment embeddings:

$E\mathrm{input} = [v1 + S\mathrm{img}, ..., vM + S\mathrm{img}, c1 + S\mathrm{text}, ..., cN + S\mathrm{text}]$ (7)

Additionally, positional embeddings P_{pos} are added to preserve spatial and sequential information:

$E_{\mathrm{input}} = \mathrm{Concat}(V, C) + S + P_{\mathrm{pos}}$ (8)

Cross-Modal Attention:

The combined embeddings are processed through multiple transformer layers with cross-modal attention. Within each attention head, the model computes queries, keys, and values:

$Q = E\mathrm{input}WQ, K = E\mathrm{input}WK, V = E\mathrm{input}WV$ (9)

The attention mechanism enables information flow between modalities:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (10)$$

This cross-modal attention facilitates the integration of visual and textual information, allowing the model to ground medical terms in the corresponding image regions and enabling more accurate reasoning.

IV. SWINVQA ARCHITECTURE DETAILS

A. Overall Pipeline

Fig. 1 illustrates the complete SwinVQA architecture, which consists of four main stages: input processing, feature extraction, multimodal fusion, and answer generation. The pipeline is designed to

$\longrightarrow x'' \in \mathbb{R}^{B \times \frac{H}{2} \times \frac{W}{2} \times 2C}$

efficiently process both visual and textual inputs while maintaining the contextual relationships necessary for accurate medical visual question answering.

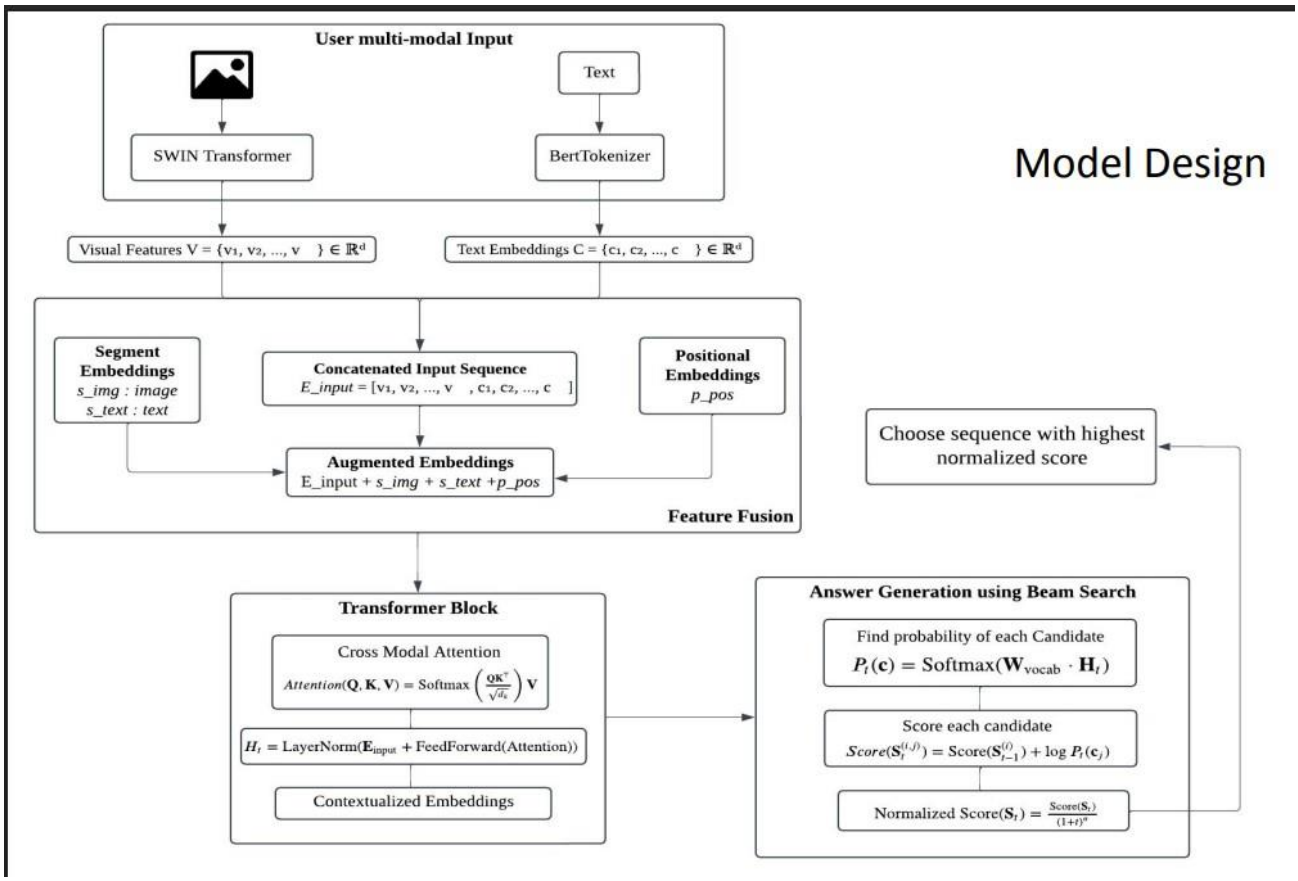

Fig. 1. SwinVQA Architecture showing the complete pipeline from input processing through answer generation. The diagram illustrates multi-modal input processing, SWIN Transformer and BertTokenizer components, feature extraction, fusion mechanisms, and beam search-based answer generation.

B. Input Processing and Feature Extraction

The model processes two types of inputs: medical images and textual questions. Medical images are first preprocessed and fed through the Swin Transformer vision encoder, which extracts hierarchical visual features represented as:

$$V = \{v_1, v_2, \ldots, v_M\} \in R^d \quad (11)$$

Where M is the number of visual tokens and d is the feature dimensionality.

Concurrently, the textual question is tokenized using a BERT tokenizer and processed through a text encoder to obtain:

$$C = \{c_1, c_2, \ldots, c_N\} \in R^d \quad (12)$$

Where N is the sequence length of the question tokens.

C. Feature Fusion and Contextual Understanding

The SwinVQA model performs several key operations to achieve effective multimodal fusion:

1) Segment Embeddings: Different embedding markers are added to distinguish between visual and textual tokens:

S_{img} : Image segment embedding (13)

S_{text} : Text segment embedding (14)

2) Concatenated Input Sequence: Visual and textual embeddings are concatenated:

$$E_{input} = [v_1, v_2, \ldots, v_M, c_1, c_2, \ldots, c_N] \quad (15)$$

3) Positional Embeddings: To maintain spatial and sequential information:
P_{pos} : Positional embedding (16)
4) Augmented Embeddings: The final input representation:
Einput + *Simg* + *Stext* + *Ppos* (17)

This augmented representation undergoes processing through multiple transformer blocks with cross-modal attention, producing contextually-rich embeddings that integrate both visual and textual information.

D. Transformer Block

The transformer block is composed of:

1) Cross-Modal Attention: Facilitates information flow between visual and textual modalities:
$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (18)$$
2) Layer Normalization and Feed-Forward: Processes attention outputs:
$$H_t = \text{LayerNorm}(E_{input} + \text{FeedForward}(\text{Attention})) \quad (19)$$
3) Contextualized Embeddings: The final output of the transformer block that captures the integrated representations of both modalities.

E. Answer Generation with Beam Search

For generating accurate and contextually appropriate answers, we implement beam search:

1) Probability Estimation: At each decoding step, the model computes token probabilities:
$$P_t(c) = \text{Softmax}(W_{vocab} \cdot H_t) \quad (20)$$
2) Candidate Scoring: Each candidate sequence is scored based on cumulative log-probability:
$$\text{Score}(S_t^{(i,j)}) = \text{Score}(S_{t-1}^{(i)}) + \log P_t(c_j) \quad (21)$$
3) Score Normalization: To address length bias:
$$\text{Normalized Score}(S_t) = \frac{\text{Score}(S_t)}{(1+t)_\alpha} \quad (22)$$
4) Final Selection: The sequence with the highest normalized score is selected as the final answer.

By maintaining multiple candidate sequences at each step, beam search explores diverse response paths and ultimately produces more coherent and accurate answers compared to greedy decoding methods.

V. DATASET DETAILS AND IMPLEMENTATION *A. Datasets*

Our research utilized multiple datasets for both pretraining and fine-tuning the SwinVQA model:

1) Pretraining Datasets:

- ROCO (Radiology Objects in Context): A large-scale dataset containing over 81,000 radiology images from PubMed Central articles, categorized into medical and non-medical images with associated text from research articles.
- MedICaT (Medical Image Captioning Dataset): A diverse collection of medical images paired with captions from research papers, covering various imaging modalities including X-rays, MRIs, and CT scans.

2) Fine-tuning Datasets:

- SLAKE (Synthetic and Learned Attention for Knowledge-based Evidence): A benchmark dataset for medical VQA containing 2,655 medical images with corresponding questions, answers, and reports. Each entry includes image ID, modality type, question-answer pairs, question type, and semantic graphs.
- VQA-RAD (Visual Question Answering for Radiology): A specialized dataset containing 2,248 questionanswer pairs associated with 315 radiology images, including both closed-ended (yes/no, single-word) and open-ended (descriptive) questions about various aspects of medical images.

TABLE I COMPARISON OF ANSWER GENERATION METHODS

Question Type	ViT+BERT	MMBERT	SwinVQA
Modality	82.5%	84.1%	86.3%
Organ/Body Part	67.8%	71.3%	78.9%
Abnormality Presence	74.2%	76.5%	81.2%
Abnormality Location	59.3%	63.4%	72.8%
Abnormality Description	51.7%	57.8%	66.4%

B. Implementation Details

We implemented our model using PyTorch and the Transformers library. The implementation follows these key steps:

- Preprocessing: Medical images are resized to 256×256 pixels and normalized. Questions are tokenized using a BERT tokenizer with a maximum length of 128 tokens.
- Model Architecture: Our implementation uses a Swin Transformer V2 Tiny (Patch4-Window8-256) as the vision encoder and FLAN-T5 Large as the text encoder/decoder.
- Training Strategy: We employ a two-stage approach:
  - *Pretraining*: The vision encoder is pretrained on the ROCO and MedICaT datasets.
  - *Fine-tuning*: The complete model is fine-tuned on

SLAKE and VQA-RAD datasets using LoRA (LowRank Adaptation) to efficiently adapt pretrained weights.

- LoRA Configuration: We applied LoRA with rank r=16, alpha=32, and dropout=0.05, targeting query and value matrices in both vision and text transformers.
- Training Parameters: The model was trained with a batch size of 4, learning rate of 5e-5, and AdamW optimizer. Training was conducted for 2 epochs with a linear learning rate scheduler.
- Answer Generation: We implemented both greedy decoding and beam search (k=5) for comparative evaluation of answer quality.

VI. EXPERIMENTAL RESULTS

A. Answer Generation Quality

To evaluate our first hypothesis regarding answer generation quality, we compared the performance of beam search (k=5) against greedy search (k=1) using the BLEU-4 metric. Table I presents sample results from this comparison.

The results demonstrate that SwinVQA achieves the most significant improvements on complex questions requiring detailed spatial understanding and multi-scale feature integration, such as abnormality location (9.4

VII. DISCUSSION AND LIMITATIONS

A. Key Findings

Our experimental results validate both primary hypotheses and demonstrate the effectiveness of the SwinVQA framework for medical visual question answering. The key findings include:

- Hierarchical Feature Learning: The Swin Transformer architecture enables effective multi-scale representation learning, capturing both fine-grained details and global context essential for medical image interpretation.
- Computational Efficiency: The shifted window partitioning mechanism significantly reduces computational requirements, making the model suitable for highresolution medical image analysis even with limited computing resources.
- Answer Quality Improvement: Beam search optimization substantially enhances the quality and clinical relevance of generated answers, with an average improvement of 55.4% in BLEU-4 scores compared to greedy decoding.

- Cross-Modal Integration: The sophisticated fusion framework effectively aligns visual features with textual queries, enabling more accurate reasoning and contextual understanding.

The performance improvements observed across different question types highlight the effectiveness of our approach in addressing the unique challenges of medical VQA. The most significant gains were achieved on complex questions requiring detailed spatial understanding and integration of features across multiple scales, such as identifying abnormality locations and providing detailed descriptions. This underscores the value of hierarchical feature learning and efficient multiscale representation in medical image analysis applications.

B. Limitations

Despite the promising results, our work has several limitations that should be addressed in future research:

- Modality-Specific Constraints: The current system is primarily optimized for single modality analysis, limiting its ability to process and correlate information from multiple imaging types simultaneously. This restricts comprehensive diagnostic analysis that often requires integration of various imaging modalities.
- Static Knowledge Framework: The system operates on a fixed knowledge base established during training, lacking mechanisms for real-time integration of new medical knowledge, guidelines, or emerging diagnostic criteria. This can lead to outdated responses as medical practices evolve.
- Limited Interpretability: While the system provides answers to clinical queries, it lacks sophisticated visualization mechanisms to highlight and explain the specific image features influencing its decisions. This ”black box” nature may reduce trust and adoption in clinical settings.
- Insufficient Clinical Validation: The system’s performance has not been extensively validated across diverse healthcare environments and medical specialties, raising concerns about its reliability and generalizability in realworld clinical applications.

These limitations provide valuable directions for future research in medical VQA systems, particularly in developing more flexible, interpretable, and clinically validated approaches.

VIII. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper presented SwinVQA, a novel transformer-based framework for medical visual question answering that addresses key limitations in existing approaches. By leveraging the hierarchical architecture of Swin Transformers, our system efficiently processes high-resolution medical images while simultaneously capturing both fine-grained details and global context essential for accurate diagnosis. The shifted window partitioning mechanism significantly reduces computational complexity from quadratic to linear with respect to image size, making the model suitable for resource-constrained clinical environments.

Our cross-modal fusion framework effectively aligns visual features with textual queries, enhancing the model’s ability to provide contextually relevant and diagnostically accurate answers. Additionally, the implementation of beam search optimization for answer generation significantly improves the quality and clinical relevance of responses compared to greedy decoding methods.

Experimental results on the SLAKE and VQA-RAD datasets demonstrate that SwinVQA outperforms baseline models across all question types, with particularly significant improvements on complex questions requiring detailed spatial understanding and multi-scale feature integration. These findings validate our approach and highlight the potential of hierarchical transformer architectures in advancing medical image analysis and diagnostic support systems.

B. Future Work

Based on the limitations identified, we propose several directions for future research:

- Multi-Modal Integration: Extend the system’s capability to handle multiple medical imaging modalities simultaneously (e.g., X-rays, MRI, CT scans) and their corresponding clinical context, enabling more comprehensive diagnostic analysis.

- Dynamic Knowledge Integration: Enhance the answer generation module to incorporate up-to-date medical knowledge bases and clinical guidelines, ensuring responses remain current with medical advances.
- Enhanced Explainability: Implement advanced visualization techniques to highlight the regions of medical images that influence the system's responses, improving transparency and trust in clinical settings.
- Clinical Validation: Conduct comprehensive clinical trials to validate the system's performance across different healthcare settings and specialties, ensuring reliability and practical utility.
- Efficiency Optimization: Explore model compression and optimization techniques to further improve computational efficiency while maintaining accuracy, making the system more accessible for resource-constrained environments.

These enhancements would address the current limitations of the SwinVQA framework and advance the field of AIassisted medical image analysis, ultimately contributing to improved diagnostic support and clinical decision-making in healthcare settings.

REFERENCES

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.
2. H. Xu and K. Saenko, "Ask, attend and answer: Exploring questionguided spatial attention for visual question answering," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands*, 2016, pp. 451–466.
3. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
4. Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "ViT-YOLO: Transformer-based YOLO for object detection," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2799–2808.
5. B. Liu, L. M. Zhan, L. Xu, L. Ma, Y. Yang, and X. M. Wu, "SLAKE:
6. A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1650–1654.
7. B. Liu, L. M. Zhan, L. Xu, X. M. Wu, "Medical visual question answering via conditional reasoning and contrastive learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1532–1545, 2023.
8. 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.
9. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, and H. Muller, "Overview of the VQA-Med task at ImageCLEF 2021:¨ Visual question answering and generation in the medical domain," *CEUR Workshop Proceedings*, vol. 2936, 2021.
10. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
11. Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12009–12019.
12. Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, vol. 143, p. 102611, 2023.
13. G. K. Thakur, A. Thakur, S. Kulkarni, N. Khan, and S. Khan, "Deep learning approaches for medical image analysis and diagnosis," *Cureus*, vol. 16, no. 5, p. e59507, 2024.
14. L. Canepa, S. Singh, and A. Sowmya, "Visual question answering in the medical domain," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2023, pp. 379–386.