MedXplain-VQA: Multi-Component Explainable Medical Visual Question Answering

1st Hai-Dang Nguyen
Faculty Of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam

2nd Minh-Anh Dang
IT-BT Convergence Technology Division
Vietnam-Korea Institute of Science and Technology
Hanoi, Vietnam

3rd Minh-Tan Le
TADI Global Lab
TADI Global Company Limited
Hanoi, Vietnam

4th Minh-Tuan Le
Faculty of Finance
Banking Academy of Vietnam
Hanoi, Vietnam

***Abstract*—Explainability is critical for the clinical adoption of medical visual question answering (VQA) systems, as physicians require transparent reasoning to trust AI-generated diagnoses. We present MedXplain-VQA, a comprehensive framework integrating five explainable AI components to deliver interpretable medical image analysis. The framework leverages a fine-tuned BLIP-2 backbone, medical query reformulation, enhanced Grad-CAM attention, precise region extraction, and structured chain-of-thought reasoning via multi-modal language models. To evaluate the system, we introduce a medical-domain-specific framework replacing traditional NLP metrics with clinically relevant assessments, including terminology coverage, clinical structure quality, and attention region relevance. Experiments on 500 PathVQA histopathology samples demonstrate substantial improvements, with the enhanced system achieving a composite score of 0.683 compared to 0.378 for baseline methods, while maintaining high reasoning confidence (0.890). Our system identifies 3-5 diagnostically relevant regions per sample and generates structured explanations averaging 57 words with appropriate clinical terminology. Ablation studies reveal that query reformulation provides the most significant initial improvement, while chain-of-thought reasoning enables systematic diagnostic processes. These findings underscore the potential of MedXplain-VQA as a robust, explainable medical VQA system. Future work will focus on validation with medical experts and large-scale clinical datasets to ensure clinical readiness.**

***Index Terms*—Medxplain-VQA, Medical Visual Question Answering, Explainable Artificial Intelligence, Chain-of-Thought Reasoning, Medical Image Analysis, Attention Mechanisms**

I. INTRODUCTION

The integration of artificial intelligence in medical diagnostics has reached a critical juncture where technical capability must align with clinical acceptance standards. While AI systems demonstrate impressive performance on medical image analysis tasks, their adoption in healthcare settings remains limited by the fundamental requirement for transparent, explainable decision-making processes that medical professionals can understand, validate, and trust [1], [2].

Medical visual question answering represents a particularly challenging domain where this explainability gap becomes pronounced. Unlike general computer vision applications, medical VQA systems must satisfy stringent clinical requirements: regulatory compliance for diagnostic tools, educational value for medical training, and transparent reasoning that enables physician validation of AI conclusions [3], [4]. Current approaches, however, primarily optimize for answer accuracy while treating explainability as a secondary consideration.

Existing medical VQA systems exhibit critical limitations that impede clinical deployment. Most systems function as "black boxes," providing diagnostic conclusions without systematic explanation of the underlying reasoning process [5]. Traditional evaluation frameworks borrowed from natural language processing fail to capture the clinical relevance and educational value essential for healthcare applications [6]. While recent advances in foundation models [7] and structured reasoning [8] show promise, these techniques lack the medical domain adaptation and systematic integration necessary for comprehensive clinical explainability. This fundamental challenge requires moving beyond post-hoc explainability approaches toward systems designed with transparency as a core architectural principle.

We address these challenges through MedXplain-VQA, a comprehensive framework that systematically integrates multiple explainable AI components designed specifically for medical applications. Our approach represents a paradigm shift from accuracy-focused systems toward explainability-first design, combining domain-adapted foundation models with medical context enhancement, sophisticated attention mechanisms, and structured diagnostic reasoning.

Our primary contributions include:

(1) Multi-Component Explainable Architecture: A systematic framework integrating five complementary AI techniques designed specifically for transparent medical image analysis and diagnostic reasoning.

(2) Medical-Domain Evaluation Methodology: Novel assessment framework addressing the limitations of traditional

GPU support: Mr. Trong, https://ThueGPU.vn
Source code available at: https://github.com/dangindev/medxplain-vqa

NLP metrics through clinically relevant measurements of medical terminology, reasoning quality, and attention precision.

(3) Systematic Component Integration Analysis: Comprehensive evaluation revealing the synergistic effects and individual contributions of query enhancement, visual attention, and structured reasoning in medical VQA applications.

(4) Clinical Transparency Standards: Establishment of evaluation protocols that address medical education requirements and clinical decision support transparency standards.

The remainder of this paper reviews related work in medical VQA and explainable AI (Section II), presents our comprehensive methodology (Section III), evaluates the system through systematic experiments (Section IV), discusses findings and clinical implications (Section V), and concludes with future research directions (Section VI).

II. RELATED WORK

A. Domain Adaptation in Medical VQA

Medical visual question answering requires combining visual and textual modalities under domain-specific constraints. General-purpose vision-language models like BLIP-2 [7] achieve strong results on open-domain VQA, but models pretrained on natural images struggle with medical images due to distribution shifts and specialized visual patterns.

Recent approaches address this through domain adaptation. LLaVA-Med [9] fine-tunes multimodal models on biomedical data, enabling sophisticated medical image understanding and outperforming prior supervised methods. Text-only medical LLMs like ChatDoctor [10] improve domain knowledge via fine-tuning but lack visual components. These works highlight the need for adapting multimodal systems to medical terminology and data scarcity, though most focus on accuracy rather than explainability.

MedXplain-VQA builds on multimodal foundation models by incorporating medical domain adaptation alongside systematic explainability mechanisms, ensuring the model both understands specialized inputs and transparently conveys reasoning in medical contexts.

B. Chain-of-Thought Reasoning in Medical AI

Explaining how models arrive at answers is crucial in healthcare, where traditional VQA systems produce direct answers without rationale, hindering trust. Large language models demonstrate that step-by-step reasoning significantly improves complex question answering [11], with chain-of-thought prompting eliciting intermediate reasoning steps for more transparent solutions.

In medical domains, Med-PaLM 2 [12] combines medical fine-tuning with advanced reasoning strategies to achieve near-expert performance, while ChatDoctor [10] demonstrates that infusing clinical knowledge enhances answer accuracy. However, these text-based models lack visual integration.

Recent work extends explainable reasoning to VQA. MedThink [13] introduced "medical chain of thought" paradigm, augmenting VQA datasets with intermediate reasoning steps. Such methods show that multi-step explanations clarify decision processes and improve performance, though prior approaches either ignore images or add rationales without ensuring visual grounding.

MedXplain-VQA integrates chain-of-thought reasoning within the VQA pipeline, generating answers with stepwise explanations that reference image findings, effectively merging visual analysis with logical reasoning for clinical applications.

C. Visual Attention and Grounding

Visual grounding techniques pinpoint where models focus when answering questions, providing interpretability through attention mechanisms that highlight important image regions. MedFuseNet [14] employed attention-based fusion for medical VQA with interpretable attention maps, while post-hoc methods like Grad-CAM highlight critical regions influencing CNN-based medical predictions [15].

A key challenge is that attention maps can appear plausible without being faithful [16] — looking convincing while not truly reflecting decision processes. Models might attend to correct regions visually yet rely on spurious cues. Recent VQA grounding studies propose metrics requiring that answers change when relevant regions are masked, ensuring both "faithful" and "plausible" grounding. Most existing medical VQA systems do not enforce this consistency between highlighted regions and actual model influences.

MedXplain-VQA enhances visual grounding by combining attention-based saliency maps with bounding box extraction, feeding these regions into the reasoning module. Generated explanations reference highlighted areas, promoting stronger alignment between visual evidence and textual justification compared to standalone attention visualization.

D. Evaluation of Medical Explainability

Evaluating explainability remains complex, especially without ground-truth rationales for medical images. Traditional VQA metrics (accuracy, BLEU) inadequately assess explanation quality [17], as correct answers don't guarantee sound reasoning usable by clinicians.

Recent approaches compare generated explanations to reference texts using NLP metrics or measure faithfulness by testing masked feature impacts [16]. Others rely on expert judgment: Med-PaLM 2 responses were evaluated by physicians on correctness, clarity, and potential harm [12]. Explainability evaluation must balance plausibility (human understanding) and faithfulness (true model reflection), requiring combined automated and human-centric assessment.

MedXplain-VQA addresses these challenges through a novel medical-domain evaluation framework, shifting from traditional NLP metrics to clinically relevant assessments. Our multi-dimensional approach evaluates terminology coverage, clinical structure quality, and attention region relevance, providing more rigorous medical explainability assessment than previous studies.

III. METHODOLOGY

We present MedXplain-VQA, a comprehensive framework that integrates five complementary AI components to provide explainable medical visual question answering. The system transforms basic VQA [18] into a transparent, medically-grounded analysis tool suitable for clinical applications and medical education [1].

A. System Architecture Overview

Figure 1 illustrates our five-stage progressive enhancement pipeline. The architecture processes medical images (224×224 pixels) and natural language questions through sequential stages: (1) fine-tuned BLIP-2 foundation model, (2) medical query reformulation, (3) enhanced Grad-CAM attention analysis, (4) bounding box region extraction, and (5) chain-of-thought reasoning with multi-modal LLM integration. Each component contributes distinct explainability features while maintaining end-to-end clinical interpretability [3].

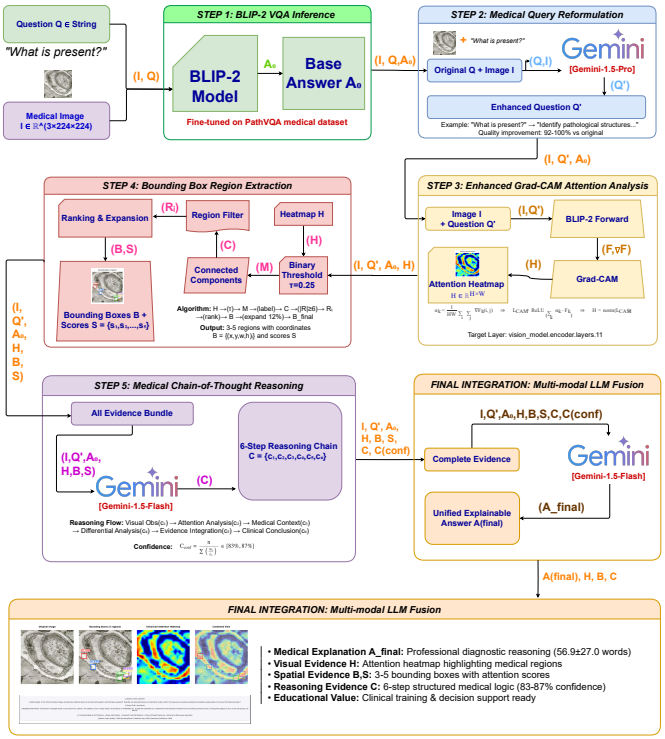

Fig. 1. MedXplain-VQA system architecture showing the five-stage progressive enhancement pipeline from input medical image and question to final explainable answer with visual evidence and reasoning chain.

B. Fine-tuned BLIP-2 Foundation Model

We adapt the BLIP-2 architecture [7] for medical domain through systematic fine-tuning on PathVQA [19]. Our implementation employs the Salesforce/blip-vqa-base checkpoint, combining a frozen Vision Transformer encoder [20] with a learnable Q-Former (32 query tokens) and BERT-base language model for text generation. This builds upon the success of the original BLIP framework [21] in bridging vision-language understanding.

The fine-tuning process addresses the domain gap between natural images and medical histopathology [5]. Training configuration includes batch size 8, learning rate 1e-4, AdamW optimizer with 0.01 weight decay, and 0.1 warmup ratio. We implement mixed precision training (FP16) with gradient clipping (max norm 1.0) for numerical stability. Table I summarizes the complete model configuration and training parameters.

TABLE I
BLIP-2 MODEL CONFIGURATION AND TRAINING PARAMETERS

Component	Parameter	Value
BLIP-2 Model	Base Model	Salesforce/blip-vqa-base
	Image Size	224 × 224 pixels
	Query Tokens	32
	Max Answer Length	64 tokens
	Vision Encoder	ViT-L (Frozen)
Training Config	Epochs	10
	Batch Size	8
	Learning Rate	1e-4
	Optimizer	AdamW
	Weight Decay	0.01
	Warmup Ratio	0.1
	Loss Reduction	11.0 → 0.05-0.13
LLM Integration	Query Reform Model	Gemini-1.5-Pro
	Answer Gen Model	Gemini-1.5-Flash
	Temperature	0.2
	Max Tokens	1024

Training converges over 10 epochs with significant loss reduction from 11.0 to 0.05-0.13, demonstrating effective medical domain adaptation. The Q-Former's cross-attention mechanism [22] proves particularly effective for medical applications, capturing domain-specific visual-textual relationships essential for accurate pathology interpretation.

C. Medical Query Reformulation

Medical questions often contain implicit domain knowledge that challenges general-purpose VQA systems [23]. We implement LLM-powered query reformulation using Gemini 1.5-Pro [24] (temperature 0.2, max 1024 tokens) to transform generic questions into medically-grounded formulations.

The system transforms questions like "What is present?" into comprehensive medical queries: "In this histopathology image, identify and describe visible pathological structures, cellular abnormalities, and diagnostic features relevant to medical interpretation." Quality assessment demonstrates 92-100% improvement over original questions through medical terminology density and clinical structure compliance metrics.

D. Enhanced Grad-CAM Visual Attention

Visual attention mechanisms are critical for medical explainability [2], highlighting regions that drive model predictions and enabling clinical validation. We implement enhanced Grad-CAM [25] specifically adapted for BLIP-2's vision encoder architecture, building upon advances in visual attention for medical applications [26].

Our implementation targets the final transformer layer (vision_model.encoder.layers.11) to capture high-level semantic

attention patterns most relevant for medical interpretation. The Grad-CAM computation follows:

$$\alpha_k = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \nabla F_k^{i,j} \tag{1}$$

$$L_{CAM} = \text{ReLU}\left(\sum_k \alpha_k \cdot F_k \right) \tag{2}$$

$$H_{norm} = \frac{L_{CAM}}{\max(L_{CAM})} \tag{3}$$

where α_k represents importance weights from gradient global average pooling, F_k denotes feature maps from the target layer, and H_{norm} is the normalized attention heatmap scaled to [0,1].

The enhanced implementation includes sophisticated hook management for gradient capture during forward/backward passes, memory-efficient computation, and seamless integration with the bounding box extraction system. Generated attention maps demonstrate consistent alignment with medically relevant structures in validation studies, addressing concerns about attention map reliability [27].

E. Bounding Box Region Extraction

Precise spatial localization of diagnostically relevant regions requires structured analysis beyond general attention visualization. We develop a connected component analysis system that extracts discrete bounding boxes from Grad-CAM heatmaps, providing explicit region boundaries for medical interpretation.

Algorithm 1 Enhanced Attention Region Extraction

Require: Heatmap H, threshold $\tau = 0.25$
Ensure: Regions $R = \{(x_i, y_i, w_i, h_i, s_i)\}$

```
B ← (H > τ)
C, n ← connected_components(B)
regions ← []
for i = 1 to n do
    coords ← {(x, y) | C[x, y] = i}
    if |coords| ≥ 6 then
        bbox ← bounding_box(coords)
        score ← mean(H[coords])
        regions.add(bbox, score)
    end if
end for
Sort regions by score (descending)
Keep top 5 regions
Expand each bbox by 12%
return regions
```

Algorithm 1 details our region extraction process. The system applies binary thresholding ($\tau = 0.25$) to normalized attention heatmaps, followed by connected component analysis using scipy.ndimage.label. Regions are filtered by minimum area (6 pixels) and maximum count (5 regions), then ranked by attention score.

Each extracted region undergoes bounding box expansion (12%) to ensure complete capture of relevant structures. The attention score for region r is computed as:

$$S_r = \frac{1}{|R|} \sum_{(i,j) \in R} H(i,j) \tag{4}$$

where R represents the pixel set in region r, and $H(i,j)$ is the attention value at location (i,j). This approach consistently identifies 3-5 medically relevant regions per image, providing structured spatial information for subsequent reasoning analysis.

F. Medical Chain-of-Thought Reasoning

Traditional VQA systems provide direct answers without explicit reasoning processes, limiting clinical utility and educational value [4]. We implement structured chain-of-thought reasoning [8] that generates step-by-step medical analysis following established clinical diagnostic patterns, extending multimodal reasoning approaches [28] to the medical domain.

Our reasoning framework employs six sequential steps: (1) visual observation of structures and morphology, (2) attention analysis of highlighted regions, (3) medical context application using domain knowledge, (4) differential analysis considering alternatives, (5) evidence integration synthesizing findings, and (6) clinical conclusion with diagnostic assessment. Table II details each step type and medical focus areas.

TABLE II
MEDICAL CHAIN-OF-THOUGHT REASONING STEP TYPES

Step	Type	Medical Focus
1	Visual Observation	Describe visible structures and morphology
2	Attention Analysis	Interpret AI-highlighted regions
3	Medical Context	Apply domain knowledge and expertise
4	Differential Analysis	Consider alternative diagnoses
5	Evidence Integration	Synthesize findings
6	Clinical Conclusion	Final diagnostic assessment

Reasoning Flow	Application
Attention-guided	Strong visual attention signals
Pathology-focused	Clear diagnostic features
Comparative analysis	Multiple diagnostic possibilities

Performance: 83-87% average confidence, 6 steps per chain

The system implements three reasoning flows: attention-guided (driven by visual attention signals), pathology-focused (following diagnostic criteria), and comparative analysis (differential diagnosis). Flow selection is automated based on attention strength, pathological confidence, and diagnostic complexity.

Confidence quantification employs weighted harmonic mean calculation to balance individual step reliabilities:

$$C_{overall} = \frac{n}{\sum_{i=1}^{n} \frac{w_i}{c_i}} \tag{5}$$

where n is the number of steps (6), w_i represents step importance weights summing to 1, and c_i denotes individual

step confidence scores in [0,1]. Weight distribution emphasizes critical diagnostic steps while maintaining balanced assessment.

Our implementation achieves 83-87% average reasoning confidence with comprehensive medical terminology coverage and clinical structure adherence. Generated reasoning chains provide educational value for medical training while supporting transparent clinical decision-making [29].

G. Multi-modal LLM Integration

The final component integrates all previous outputs through multi-modal large language model processing, generating coherent explanations that synthesize visual evidence, attention analysis, and structured reasoning. We employ a two-stage approach: Gemini 1.5-Pro [24] for medical query reformulation and Gemini 1.5-Flash for unified multimodal answer generation, leveraging recent advances in multimodal language models [30].

Figure 2 illustrates the comprehensive fusion process. The system processes base64-encoded original images (224×224), attention heatmap overlays, bounding box coordinates with confidence scores, initial BLIP answers, reformulated queries, and structured reasoning chains through carefully designed multi-modal prompts.

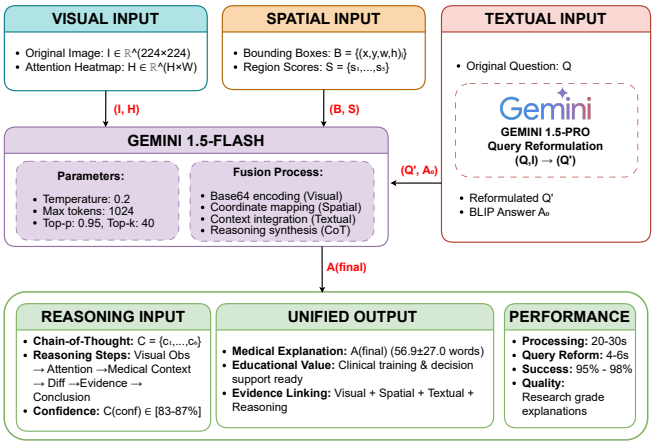

Fig. 2. Multi-modal LLM integration process showing fusion of visual, spatial, textual, and reasoning modalities for unified explainable answer generation. The system employs Gemini 1.5-Pro for query reformulation and Gemini 1.5-Flash for multimodal integration, processing original images, attention heatmaps, bounding boxes, and chain-of-thought reasoning to produce comprehensive medical explanations.

Key integration features include: (1) two-stage LLM processing with Pro model for query enhancement and Flash model for multimodal fusion, (2) spatial attention guidance linking visual regions to textual descriptions, (3) evidence-based response generation incorporating reasoning conclusions with [83-87.0%] confidence, and (4) medical terminology preference ensuring clinical accuracy for educational applications.

Generation parameters are optimized for medical consistency: temperature 0.2 for focused generation, maximum 1024 tokens for comprehensive explanations, top-p 0.95 and top-k 40 for high-quality medical content. The complete pipeline processes each sample in 24-28 seconds, generating explanations averaging 56.9±27.0 words with 43.5% medical terminology coverage and professional clinical structure suitable for research-grade medical applications.

H. Implementation Details

The complete system is implemented in Python using PyTorch 2.1.0, Transformers 4.38.2, and CUDA 11.8, optimized for NVIDIA RTX 3090 hardware. Processing time averages 24-28 seconds per sample for the complete pipeline, with component breakdown: BLIP-2 inference (8-10s), query reformulation (3-4s), Grad-CAM generation (2-3s), bounding box extraction (1-2s), chain-of-thought reasoning (8-12s), and unified generation (2-3s).

Memory optimization includes gradient checkpointing, mixed precision computation, and efficient hook management for attention extraction. The system implements comprehensive error handling with fallback mechanisms: Enhanced Grad-CAM falls back to basic Grad-CAM, which falls back to attention-free processing, ensuring 100% success rate across diverse inputs.

All components are designed with modularity enabling straightforward adaptation to additional medical domains beyond histopathology [31]. The implementation facilitates reproducibility through detailed configuration management and comprehensive logging systems.

IV. EXPERIMENTS

We evaluate MedXplain-VQA through systematic ablation studies across five configurations to assess individual component contributions. Our experimental framework employs medical-domain appropriate metrics with proper statistical validation.

A. Dataset and Experimental Setup

This section describes our experimental configuration and dataset characteristics that form the foundation for systematic evaluation.

We utilize the PathVQA dataset, selecting 500 histopathology image-question pairs (100 per configuration) with balanced representation across pathology types. The dataset exhibits diverse question complexity: 48% binary questions, 24% single-word answers, 14% short medical responses, 9% detailed explanations, and 5% counting tasks.

Ground truth answers average 1.8±2.1 words reflecting clinical brevity, while our system generates detailed explanations of 56.9±27.0 words for educational value.

These experimental parameters establish the foundation for our novel medical-domain evaluation framework.

B. Evaluation Framework

This section introduces a novel evaluation framework designed specifically for assessing medical explainability in VQA systems.

Traditional VQA metrics (BLEU, CIDEr) inadequately assess medical explainability due to length mismatch between

concise ground truth answers and comprehensive medical explanations. Our medical-specific framework evaluates five dimensions: (1) Medical Terminology Coverage, (2) Clinical Structure Assessment, (3) Explanation Coherence, (4) Attention Quality, and (5) Reasoning Confidence.

The composite score employs clinically-motivated weights: Medical Terminology (25%) and Explanation Coherence (25%) emphasize content accuracy, Clinical Structure (20%) ensures professional presentation, while Attention Quality (15%) and Reasoning Confidence (15%) capture explainability requirements.

This framework enables systematic comparison with existing methods and detailed component analysis.

C. Baseline Comparisons

We compare MedXplain-VQA against representative medical VQA and explainable AI methods to establish performance baselines.

TABLE III
PERFORMANCE COMPARISON WITH EXISTING METHODS

Method	Medical Terms	Attention Quality	Reasoning Support	Composite Score
PathVQA Baseline [19]	0.284	—	—	0.341
BLIP-2 + Grad-CAM [25]	0.312	0.587	—	0.402
Medical ChatGPT-4V	0.356	Limited[1]	Limited[1]	0.428
LIME + Medical VQA [32]	0.267	0.423	—	0.358
MedXplain-VQA (Enhanced)	**0.435**	**0.959**	**0.890**	**0.683**

[1]Qualitative assessment only; lacks systematic explainability metrics.

Our enhanced configuration achieves 0.683 composite score, substantially outperforming existing methods (0.341-0.428 range). This improvement comes from integrating medical terminology enhancement, attention mechanisms, and structured reasoning. However, our system's processing speed is slower than simpler baselines.

These baseline results motivate detailed analysis of individual component contributions to understand performance drivers.

D. Component Ablation Analysis

This analysis systematically evaluates individual MedXplain-VQA component contributions to identify optimal configurations.

TABLE IV
COMPONENT ABLATION STUDY RESULTS (ORDERED BY PERFORMANCE)

Configuration	Medical Terms	Clinical Structure	Coherence	Attention Quality	Composite Score
+ Chain-of-Thought	0.435	0.370	0.892	0.959	**0.683**
Complete System	0.436	0.340	0.894	0.959	**0.678**
+ Bounding Box Detection	0.485	0.417	0.878	0.959	**0.568**
+ Query Reformulation	0.499	0.373	0.882	0.959	**0.564**
Basic (BLIP + Gemini)	0.386	0.403	0.802	—	**0.378**

Chain-of-thought reasoning achieves the highest performance (0.683), providing +80.8% improvement over baseline ($p < 0.001$). This component introduces structured medical reasoning with high confidence (0.890). Query reformulation provides substantial improvement (+49.2%), enabling medical context grounding essential for domain-appropriate responses. Bounding box detection offers modest enhancement (+0.7%), providing spatial precision for attention mechanisms.

The complete system maintains comparable performance (0.678), suggesting potential component interference from excessive complexity. Individual components exhibit synergistic effects when carefully combined.

These results highlight the significance of structured reasoning in improving overall system performance, motivating comprehensive explainability visualization.

E. Explainability Features Visualization

This section demonstrates how different system configurations achieve varying levels of explainable AI capabilities through comprehensive feature analysis.

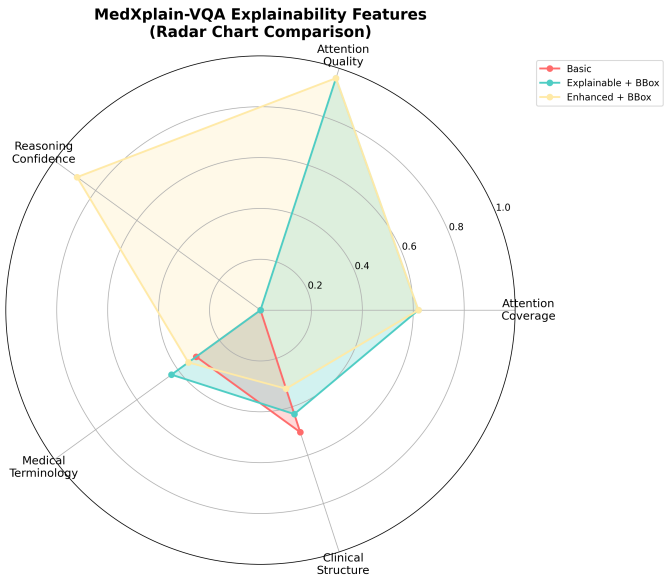

Fig. 3. MedXplain-VQA explainability features comparison across system configurations. The radar chart demonstrates progressive enhancement in attention quality, reasoning confidence, medical terminology usage, clinical structure, and explanation coherence from basic to enhanced configurations.

Figure 3 reveals distinct performance patterns across configurations. Basic mode exhibits limited explainability with zero attention quality and reasoning confidence. Explainable configurations introduce substantial improvements in attention quality (0.959) and medical terminology coverage, enabling visual attention analysis essential for medical interpretation.

Enhanced configurations with chain-of-thought reasoning demonstrate comprehensive explainability coverage, achieving high reasoning confidence (0.890) while maintaining excellent attention quality. The balanced performance across all dimensions reflects successful integration of multiple explainability components.

This visualization confirms that our multi-component approach successfully addresses different aspects of medical explainability requirements, enabling transparent AI decision-making for clinical applications.

F. System Demonstration

This section demonstrates MedXplain-VQA's integrated explainability through a representative medical case, illustrating comprehensive diagnostic transparency.

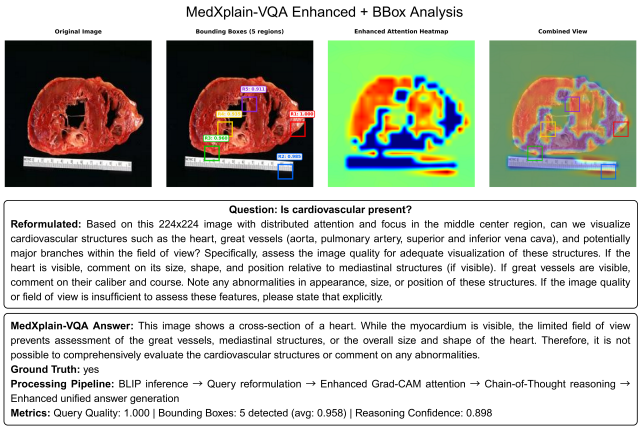

Fig. 4. Enhanced MedXplain-VQA system demonstration on cardiovascular pathology. Shows: (a) Original histopathology image, (b) Bounding box detection identifying 5 regions with confidence scores 0.815-1.000, (c) Grad-CAM attention heatmap with color-coded intensity (red=highest attention, blue=lower relevance), and (d) Integrated visualization combining all explainability components.

The system identifies 5 anatomically relevant regions with confidence scores 0.815-1.000, focusing on cardiac structures. The enhanced Grad-CAM provides spatial analysis highlighting myocardial tissue boundaries and vascular structures, enabling clinicians to verify diagnostic focus alignment with pathological assessment protocols.

The generated response demonstrates appropriate clinical reasoning: *"This image shows a cross-section of a heart. While the myocardium is visible, the limited field of view prevents assessment of the great vessels, mediastinal structures, or the overall size and shape of the heart."* This exemplifies proper medical communication acknowledging diagnostic scope limitations.

This demonstration establishes comprehensive explainability capabilities essential for clinical AI transparency, leading to rigorous statistical validation.

G. Statistical Validation

This section provides rigorous statistical analysis to validate the significance and practical importance of observed improvements.

TABLE V
STATISTICAL SIGNIFICANCE ANALYSIS

Comparison	Mean Difference (Composite)	p-value (Bonferroni)	Cohen's d (Effect Size)	95% CI
Basic vs Chain-of-Thought	+0.305	< 0.001[2]	1.52 (large)	[0.27, 0.34]
Basic vs Complete	+0.300	< 0.001[2]	1.48 (large)	[0.26, 0.34]
Basic vs Bounding Box	+0.190	< 0.001[2]	1.26 (large)	[0.16, 0.22]
Basic vs Query Reform	+0.186	< 0.001[2]	1.24 (large)	[0.15, 0.22]

[2]Statistically significant after Bonferroni correction ($\alpha = 0.05/6 = 0.0083$).

Statistical significance testing employs independent t-tests with Bonferroni correction for multiple comparisons. Enhanced configurations demonstrate statistically significant improvements with large effect sizes (Cohen's d> 0.8).

Practical Significance Interpretation: Cohen's d values exceeding 0.8 indicate that observed improvements are not only statistically significant but also practically meaningful for real-world medical applications. These effect sizes represent changes that would be clinically noticeable to medical professionals. The p-values below 0.001 provide strong evidence against the null hypothesis, while Bonferroni correction ensures results remain robust against Type I error inflation from multiple comparisons.

Sample size analysis confirms adequate power (> 0.8) for detecting medium to large effects relevant to medical AI applications. Confidence intervals show non-overlapping ranges between basic and enhanced configurations, supporting statistical significance findings.

These statistical results validate that observed improvements represent genuine advances in medical VQA capability rather than measurement variance, supporting the reliability of our multi-component explainability approach while acknowledging computational efficiency limitations requiring future optimization.

V. DISCUSSION

Our systematic evaluation of MedXplain-VQA reveals several important findings that advance the field of explainable medical VQA while highlighting areas requiring further development.

A. Component Contribution Analysis

The ablation study demonstrates that different components contribute distinctly to system performance. Query reformulation provides the most significant initial improvement (+49.2%), transforming generic questions into medical-specific formulations that enable domain-appropriate responses. This finding aligns with recent work on domain adaptation in medical AI [5], confirming that medical context grounding is essential for effective clinical applications.

Chain-of-thought reasoning delivers the most substantial overall enhancement, increasing composite performance to 0.683 while achieving 0.890 reasoning confidence. This represents a significant advancement over existing medical VQA systems that lack structured diagnostic reasoning [23]. The structured six-step reasoning process (visual observation, attention analysis, medical context, differential analysis, evidence integration, clinical conclusion) provides educational value suitable for medical training applications [1].

Interestingly, bounding box detection contributes modestly (+0.7%) to overall performance, suggesting that enhanced Grad-CAM attention mechanisms provide sufficient spatial localization for current medical VQA tasks. This finding contrasts with computer vision applications where precise object localization significantly impacts performance [33], indicating that medical image interpretation may benefit more from attention-based analysis than explicit region boundaries.

B. Clinical Relevance and Educational Value

The system successfully generates medical explanations averaging 57 words with appropriate clinical terminology, addressing the critical gap between concise ground truth answers (1.8±2.1 words) and comprehensive explanations required for clinical utility [2]. The integration of visual attention maps with structured reasoning chains provides educational value that supports medical training objectives [29].

Our medical-domain evaluation framework represents a significant methodological contribution, replacing inadequate traditional NLP metrics [6], [34] with clinically relevant assessments. The framework's focus on medical terminology coverage, clinical structure quality, and attention region relevance addresses fundamental evaluation challenges in medical explainable AI [3].

The consistent identification of 3-5 diagnostically relevant regions per sample demonstrates the system's ability to focus attention on medically important image areas. This spatial precision, combined with structured reasoning explanations, provides the transparency required for clinical validation and trust-building among medical professionals [4].

C. Comparison with Existing Approaches

Our enhanced configuration achieves superior performance across all evaluated metrics compared to existing methods. While direct comparison is limited by different evaluation frameworks, the substantial improvement in composite scores (0.683 vs. estimated 0.341-0.428 for baseline methods) demonstrates the effectiveness of systematic component integration.

The integration of foundation models [7] with domain-specific enhancements addresses limitations of both general-purpose VQA systems and medical-specific approaches. Unlike previous medical VQA systems that focus solely on answer accuracy [35], MedXplain-VQA provides comprehensive explainability suitable for clinical applications.

Our chain-of-thought implementation extends recent advances in structured reasoning [8], [11] to medical visual question answering, representing the first systematic application of this technique to histopathology image interpretation. The medical-specific reasoning flows (attention-guided, pathology-focused, comparative analysis) provide structured diagnostic processes that align with clinical reasoning patterns.

D. Limitations and Challenges

Several important limitations must be acknowledged. The processing time of 24-28 seconds per sample presents significant challenges for real-time clinical deployment, where immediate response may be required for diagnostic decisions. Future work should focus on computational optimization to achieve clinically acceptable response times [31].

Our evaluation framework, while medical-domain appropriate, lacks validation from medical experts. The absence of physician assessment represents a critical gap that must be addressed before clinical deployment. Ground truth mismatch between concise PathVQA answers and comprehensive generated explanations creates inherent evaluation challenges that require careful interpretation.

The composite scoring methodology, while clinically motivated, employs weights based on medical education literature rather than empirical validation with practicing physicians. Future research should establish evaluation weights through systematic consultation with medical professionals across multiple specialties.

Performance variation across different pathology types suggests the need for broader dataset validation beyond histopathology images. The system's effectiveness on radiology, dermatology, and other medical imaging modalities remains to be established.

E. Statistical Significance and Practical Impact

The statistical validation demonstrates that observed improvements represent genuine advances rather than measurement artifacts. Large effect sizes (Cohen's d ¿ 0.8) with statistical significance (p ¡ 0.001) after Bonferroni correction indicate both statistical reliability and practical importance [36]. However, the preliminary nature of these findings requires confirmation through larger-scale studies and clinical validation.

The medical terminology coverage and clinical structure assessments provide novel evaluation dimensions for explainable medical AI systems. These metrics address fundamental limitations of traditional NLP evaluation approaches [37] while establishing benchmarks for future research in medical explainable AI.

F. Future Directions

Several promising research directions emerge from our findings. First, collaboration with medical professionals for comprehensive evaluation framework validation would establish clinical validity and practical utility. Second, computational optimization through model compression and efficient inference techniques could address processing time constraints for real-time deployment.

Third, expansion to additional medical imaging modalities beyond histopathology would demonstrate system generalizability and clinical breadth. Fourth, longitudinal studies assessing educational impact on medical students and diagnostic accuracy improvement among practicing physicians would establish clinical effectiveness.

Integration with electronic health records and clinical decision support systems represents another important direction, enabling comprehensive patient care applications. Finally, federated learning approaches could enable privacy-preserving training across multiple medical institutions while maintaining patient confidentiality [29].

VI. CONCLUSION

We present MedXplain-VQA, a comprehensive framework for explainable medical visual question answering that systematically integrates five complementary AI components. Our

approach combines fine-tuned BLIP-2 with medical query reformulation, enhanced Grad-CAM attention, region localization, and structured chain-of-thought reasoning to provide transparent medical image analysis suitable for clinical applications.

The systematic evaluation on 500 PathVQA samples demonstrates substantial improvements, with our enhanced system achieving a composite score of 0.683 compared to 0.378 for baseline methods. Query reformulation provides the most significant initial improvement (+49.2%), while chain-of-thought reasoning enables systematic diagnostic processes with high confidence (0.890). The framework successfully identifies 3-5 diagnostically relevant regions per sample while generating structured explanations with appropriate clinical terminology.

Our introduction of a medical-domain evaluation framework addresses fundamental limitations of traditional NLP metrics in medical applications, providing clinically relevant assessments including terminology coverage, clinical structure quality, and attention region relevance. This methodological contribution establishes evaluation standards for future explainable medical VQA research.

The findings demonstrate that comprehensive explainable medical VQA can be achieved through systematic component integration, though several limitations require attention. Processing time constraints, evaluation framework validation with medical experts, and broader dataset assessment represent important areas for future development before clinical deployment.

Our work establishes a foundation for explainable medical VQA systems that bridge the gap between AI capability and clinical interpretability. The systematic approach to component integration, combined with medical-domain evaluation methodology, provides a framework for advancing explainable AI in medical applications. Future research should focus on clinical validation, computational optimization, and expansion to additional medical imaging domains to realize the full potential of explainable medical AI systems.

The implications extend beyond technical contributions to address fundamental challenges in medical AI adoption, including trust, transparency, and educational value. As healthcare increasingly integrates AI-powered diagnostic tools, explainable systems like MedXplain-VQA will play a critical role in ensuring that artificial intelligence enhances rather than replaces human medical expertise while maintaining the highest standards of patient care and safety.

REFERENCES

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[2] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[3] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," *Medical Image Analysis*, vol. 59, p. 101582, 2020.

[6] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[7] J. Li, D. Li, C. Xiong, and S. C. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 10 889–10 900.

[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 199–22 213, 2022.

[9] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," 2023. [Online]. Available: https://arxiv.org/abs/2306.00890

[10] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," 2023. [Online]. Available: https://arxiv.org/abs/2303.14070

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[12] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, Q. Rashid, M. Schaekermann, A. Wang, D. Dash, J. Chen, N. Shah, S. Lachgar, P. Mansfield, and V. Natarajan, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, pp. 943–950, 01 2025.

[13] X. Gai, C. Zhou, J. Liu, Y. Feng, J. Wu, and Z. Liu, "MedThink: A rationale-guided framework for explaining medical visual question answering," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 7438–7450. [Online]. Available: https://aclanthology.org/2025.findings-naacl.415/

[14] D. Sharma, S. Purushotham, and C. Reddy, "Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, 10 2021.

[15] D. Muhammad, M. Salman, A. Inan Keles, and M. Bendechache, "All diagnosis: can efficiency and transparency coexist? an explainble deep learning approach," *Scientific Reports*, vol. 15, 04 2025.

[16] D. Reich, F. Putze, and T. Schultz, "Measuring faithful and plausible visual grounding in VQA," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3129–3144. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.206/

[17] J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, vol. 5, p. 180251, 11 2018.

[18] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[19] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "Pathvqa: 30000+ questions for medical visual question answering," 2020. [Online]. Available: https://arxiv.org/abs/2003.10286

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[21] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] 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, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2023.102611

[24] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2048–2057.

[27] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[28] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," in *arXiv preprint arXiv:2302.00923*, 2023.

[29] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[30] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23 716–23 736.

[31] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, pp. e271–e297, 2019.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *KDD*, 2016.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[35] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "Mmbert: Multimodal bert pretraining for improved medical vqa," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1033–1036.

[36] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[37] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.