Fahmy K, Zorkany M, Ammar AE. LoRA-MedSim: An Enhanced Multimodal Framework for Clinical Reasoning and Realistic Patient-Doctor Interaction. *Health Psychol Res*. 2025;13(3):e81240023.
https://doi.org/10.14440/hpr.0175

Research Article

LoRA-MedSim: An Enhanced Multimodal Framework for Clinical Reasoning and Realistic Patient-Doctor Interaction

Khadeja Fahmy[1]*, Mohamed Zorkany[2], Abd El-Hady Ammar[1]

[1]Department of Communication and Electronics, Faculty of Engineering, Al-Azhar University, Nasr City, Cairo 11765, Egypt

[2]Department of Communication and Electronics, National Telecommunication Institute, Nasr City, Cairo 11765, Egypt

Keywords: Healthcare system, Artificial intelligence, Low-rank adaptation, Fine-tuning, Large language model, Visual question answering

Health Psychology Research

Vol. 13, 2025

Background

Present-day medical artificial intelligence (AI) models for medical visual question answering and report generation frequently function passively. Rather than simulating the dynamic nature of clinical workflow, they primarily respond to direct commands. This restricts their applicability in real-world healthcare settings, where complex reasoning and multi-turn dialogue are crucial.

Objective

This paper proposes a novel multimodal framework, based on Low-Rank Adaptation (LoRA), that simulates authentic doctor–patient interactions. The objective is to develop an AI assistant capable of participating in multi-turn diagnostic conversations with improved reasoning, interpreting radiological images, and addressing patient queries.

Methods

We used a curated and enriched version of the VQA-Med CLEF 2019 dataset to perform LoRA fine-tuning on a large-scale vision–language model. To train the model for both visual diagnosis and natural language interaction, the dataset was supplemented with simulated patient queries, radiological reports, and follow-up questions.

Results

Compared with baseline and prompt-based methods, our model demonstrated superior performance. It achieved higher accuracy and lower loss values while producing outputs that were more interpretable across both textual and visual domains. The model's capacity to manage intricate, multi-turn diagnostic queries was further validated through structured assessments.

Conclusion

The proposed framework represents a significant advancement toward AI assistants designed for therapeutic settings. By integrating dialogue-based reasoning with multimodal understanding, it bridges the gap between passive AI tools and active diagnostic agents capable of interacting with patient data in real-world clinical scenarios.

1. INTRODUCTION

Artificial intelligence (AI) has revolutionized healthcare in recent years by enabling diagnostic procedures that are faster, more precise, and more cost-effective. AI technologies are increasingly applied in medical imaging, clinical decision support, and predictive analytics to help physicians analyze complex data and improve patient outcomes.[1-3] With the advent of AI, clinical diagnostics is undergoing rapid transformation, particularly in medical decision support and radiography. A notable development in this field is the integration of visual and textual knowledge through vision-language models (VLMs). By combining radiological images with related clinical narratives, VLMs can offer context-aware interpretations that approximate aspects of human diagnostic reasoning.[4]

***Corresponding author:**
Khadeja Fahmy
Department of Communication and Electronics, Faculty of Engineering, Al-Azhar University, Nasr City, Cairo 11765, Egypt
Email: khadega.al_sayed@yahoo.com

Despite this potential, contemporary AI systems frequently struggle to support genuine multi-turn interactions with patients and doctors. Unlike the dynamic nature of real-world clinical consultations, most models rely on static prompts and lack adaptive, conversational reasoning.[5,6] Bridging this gap is essential for developing AI assistants that function not only as proficient communicators but also as accurate diagnosticians. Applications that enable models to interpret medical images and generate language-based outputs to assist clinicians in diagnosis and treatment planning include automated report generation, image captioning, and medical visual question answering (Med-VQA). These represent some of the most significant clinical applications of multimodal AI. Recent multimodal extensions of large language models (LLMs) and VLMs, such as BioViL and CheXzero, have shown the ability to integrate textual and visual modalities to enhance clinical comprehension. By extracting semantically rich representations from magnetic resonance imaging (MRI), computed tomography (CT) scans, and X-rays, these models facilitate clinical reasoning, anomaly identification, and differential diagnosis with minimal human input. This paradigm shift has been predominantly driven by the availability of extensive medical image datasets (e.g., MIMIC-CXR, VQA-RAD, and VQA-Med 2019) and advances in transformer-based architectures that enable multimodal fusion.[7] Nevertheless, the majority of existing systems function in constrained environments: they respond only to discrete prompts and are unable to replicate authentic doctor–patient exchanges or maintain context across multi-turn medical conversations. Current Med-VQA and report-generation systems therefore remain limited in realism, interactivity, and clinical utility, despite notable advances in multimodal learning. Most models treat clinical reasoning as a static, single-turn task in which the model receives an image and a question once, then outputs an isolated response. Such approaches fail to reflect the dynamic characters of medical consultations, where doctors pose follow-up questions, integrate multiple data (e.g., imaging and laboratory reports), and provide context-dependent responses. Furthermore, most prior work relies on prompt-based approaches with frozen VLMs using zero-shot or few-shot inference. While flexible, these methods frequently produce hallucinated outputs, suffer from limited precision, and lack applicability to intricate multi-turn scenarios. Consequently, they are ill-suited for high-stakes clinical settings where traceability and factual accuracy are crucial.[8] A more patient-centered, interactive, and adaptive modeling approach is therefore urgently needed.

In the present study, we propose a novel multimodal framework in which a VLM fine-tuned with Low-Rank Adaptation (LoRA) simulates realistic doctor–patient conversations. Unlike previous methods that passively respond to prompts, our system actively participates in multi-turn interactions by extracting structured information from laboratory results and radiological images while responding to patient-initiated queries in a conversational flow. By integrating visual perception, clinical knowledge, and dialogue reasoning, the model maintains contextual awareness throughout the consultation.

Our approach fine-tunes a pre-trained multimodal LLM with LoRA on an enriched version of the VQA-Med CLEF 2019 dataset, which we augmented with simulated doctor–patient conversations, radiological reports, and follow-up questions. This strategy preserves parameter efficiency while enabling the model to adapt to clinically relevant tasks such as radiological report generation, abnormality detection, and Med-VQA.[9,10] By combining conversational simulation with multimodal comprehension, our method provides an important step toward deployable clinical AI agents capable of participating in naturalistic, patient-centered diagnostic sessions.

Specifically, we present a LoRA-based fine-tuned multimodal framework designed to replicate full-scale, multi-turn doctor–patient conversations. The system unifies image interpretation, laboratory report comprehension, patient query handling, and radiological report generation within a single conversational pipeline, thereby extending the scope of standard Med-VQA tasks. To emulate real clinical interactions, we introduce a simulated doctor–patient dialogue environment in which the model must reason over multimodal inputs and respond to sequential patient queries. Compared to prompt-based and zero-shot approaches, our framework achieves significant enhancements in clinical accuracy, response consistency, and contextual understanding. Evaluated on the VQA-Med CLEF 2019 dataset enriched with simulated patient queries, the model demonstrates exceptional performance across both language and visual tasks, achieving state-of-the-art accuracy and loss metrics.

2. RELATED WORK

A crucial criterion for assessing multimodal models in clinical imaging interpretation is Med-VQA. Early techniques relied on hand-crafted pipelines or convolutional neural network–recurrent neural network architectures, whereas later techniques employed pretrained VLMs such as MMBERT,[11] TUA1,[12] and WSDAN.[13] The most widely used dataset in this domain remains VQA-Med 2019.[14]

Despite progress, current models are still limited. Most rely on prompt-based learning, lack contextual interaction, and function in a single-turn environment. Additionally, Med-VQA techniques typically address isolated question answering tasks rather than more complex responsibilities such as radiological report generation or interactive patient communication. With the advent of LLMs and conversational agents, research has increasingly turned toward simulating multi-turn medical discussions. Frameworks such as Agent Clinic[15] and Med-PMC[16] introduce interactive evaluation setups in which a model assumes the role of a patient agent being questioned by a doctor. These systems aim to assess thinking and decision-making processes by simulating diagnostic dialogues. However, most approaches are unidirectional; the model acts as the questioner while the simulated patient merely responds, with limited support for bidirectional flow or patient-initiated queries.

Parameter-efficient fine-tuning techniques such as LoRA enable domain-specific customization of LLMs without full retraining.[9] Recent developments in medical vision-language pretraining have further explored strategies for enhancing clinical performance and representation quality. For example, Liu *et al.*[17] proposed G2D, a global-to-local training approach for dense radiography representation using paired text–image data to improve localization and semantic alignment. Qin *et al.*[18] introduced Adaptor, a contrastive learning method with frozen visual encoders that enables parameter-efficient training across a variety of medical applications. Similarly, IMITATE, proposed by Shah *et al.*,[19] guided vision–language pretraining on medical datasets under organized supervision, hierarchically leveraging clinical prior knowledge. Although these methods focused on large-scale pretraining to extract robust and generalizable visual–textual features, our approach instead adapts a general-purpose multimodal foundation model for a specific clinical workflow—namely, modeling multi-turn doctor–patient interactions and visual question answering (VQA) tasks—through LoRA-based fine-tuning. This paradigm supports efficient adaptability with minimal computational overhead, making it more deployable in real-world clinical settings.

LoRA has already been applied in medical contexts, as demonstrated by Med-Flamingo[10] and Qwen-Med[20] to enable lightweight tailoring of multimodal models for knowledge retrieval and question answering. Additional studies also highlight its utility for medical adaptation.[6,9,11,13] However, despite outperforming zero-shot prompting, these methods still face significant drawbacks. First, most concentrate on single-turn interactions and fail to capture the intricacy of real-world diagnostic conversations, where contextual reasoning and follow-up questions are crucial. Second, they are often evaluated on isolated subtasks (e.g., image-based question answering) rather than within fully replicated diagnostic environments that integrate imaging, structured laboratory data, and free-text clinical discourse. Third, they lack adaptive reasoning mechanisms, reducing their ability to dynamically adjust responses to evolving patient input. Our proposed LoRA-MedSim addresses these deficiencies by facilitating multi-turn, multimodal consultation. It combines conversational reasoning with vision–language interpretation to replicate authentic doctor–patient interactions in a single, validated pipeline. Specifically:

- Integration with clinical communication: our framework embeds Med-VQA within realistic clinical simulations, enabling multi-turn question answering grounded in multimodal data, such as laboratory data and X-ray images
- Patient-centered perspective: Unlike prior work, our system simulates clinical consultations from the patient's point of view. Patients can initiate inquiries about their condition (e.g., test results, anomalies), and the model responds with clinically grounded information
- Task-specific adaptation: A VLM is fine-tuned using LoRA to perform both image-grounded responses and radiology report generation, seamlessly embedded within a simulated dialogue environment.

3. SUGGESTED CRITERIA

The proposed system employs a LoRA-fine-tuned Qwen2-VLM model to interpret multimodal clinical inputs, including laboratory results and radiology images. By integrating visual elements, structured data, and conversational history, the system supports realistic multi-turn doctor–patient interactions. As a result, the diagnostic reasoning process is replicated, and context-aware, patient-centered responses are generated across multiple consultation rounds.

3.1. SYSTEM OVERVIEW

We present a novel multimodal framework that integrates natural language interaction with image-grounded medical reasoning to replicate authentic doctor–patient encounters (Figure 1). The system is based on Qwen2-VLM, a large VLM fine-tuned with LoRA on a clinically enhanced version of the VQA-Med CLEF 2019 dataset.

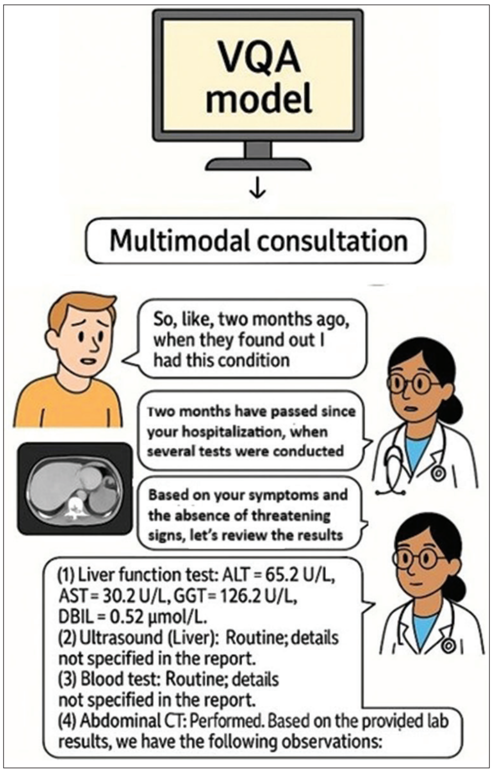
VQA model
Multimodal consultation
So, like, two months ago, when they found out I had this condition
Two months have passed since your hospitalization, when several tests were conducted
Based on your symptoms and the absence of threatening signs, let's review the results
(1) Liver function test: ALT = 65.2 U/L, AST = 30.2 U/L, GGT = 126.2 U/L, DBIL = 0.52 μmol/L.
(2) Ultrasound (Liver): Routine; details not specified in the report.
(3) Blood test: Routine; details not specified in the report.
(4) Abdominal CT: Performed. Based on the provided lab results, we have the following observations:

Figure 1. Overview of the proposed system. The model simulates a real clinical consultation by interpreting multimodal inputs (e.g., X-rays, laboratory reports), maintaining dialogue context, and generating patient-facing answers and diagnostic summaries across multiple turns.

Abbreviations: ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; CT: Computed tomography; DBIL: Direct bilirubin; GGT: Gamma-glutamyl transferase; VQA: Visual question answering.

The model engages in multi-turn dialogues with a virtual patient while processing visual and structured data, such as radiological images and laboratory reports. Through this interaction, it can extract findings, answer questions, and generate summaries that mirror real clinical consultations.

This framework operates through three fundamental steps: (i) perceiving multimodal inputs; (ii) comprehending and reasoning contextually; and (iii) generating interactive, patient-aware replies. Over several rounds, this framework enables the simulation of human-like diagnostic conversations.

3.2. MULTIMODAL INPUT PROCESSING

The system is designed to process diverse clinical inputs:

- A pre-trained vision transformer encodes radiology images (e.g., X-rays, MRI) via the Qwen2-VLM visual backbone
- Laboratory reports are converted into key–value pairs (e.g., alanine aminotransferase [ALT] = 65.2 U/L) and incorporated into the input prompt in structured text format
- Optional metadata (e.g., patient age or symptom descriptions) may be included to enhance reasoning performance.

All inputs are projected into a shared multimodal embedding space, allowing the model to reason collaboratively across different data modalities.

3.3. LORA FINE-TUNING PROCEDURE

Qwen2-VLM was specialized for clinical reasoning and interaction using an expanded version of the VQA-Med CLEF 2019 dataset with LoRA-based fine-tuning. In addition to image–question–answer triplets, the enhanced dataset incorporates synthetic multi-turn patient dialogues grounded in laboratory and imaging results.

LoRA was applied to the transformer decoder's attention layers with rank $r = 8$ and $\alpha = 32$, optimizing a limited set of parameters while preserving the full capability of the base model. After training, the LoRa adaptor was merged with the base model to create a single deployable checkpoint.

Training was performed on an NVIDIA A100 GPU using a batch size of eight and a learning rate of 5×10^{-5}, and mixed precision for five epochs. Cross-entropy loss was used for both answer generation and follow-up response prediction.

3.4. MULTI-TURN DOCTOR–PATIENT INTERACTION SIMULATION

In contrast to prompt-based systems, our approach incorporates conversational state tracking, enabling the simulated patient and the model-as-doctor to engage in back-and-forth conversations. Each consultation begins with a patient's concern or symptom. Using image and laboratory data, the model then generates a sequence of clarifying questions and answers.

The model is capable of:

- Analyzing visual indicators (e.g., "no signs of cardiomegaly")
- Obtaining and interpreting test results (e.g., "elevated ALT suggests liver dysfunction")
- Answering follow-up inquiries from the patient
- Producing clinical findings using standard medical terminology.

This simulation environment not only increases real-world applicability but also enables a more thorough assessment of multimodal knowledge across multiple stages of reasoning.

4. EXPERIMENTS AND RESULTS

The efficacy of the proposed method was assessed through extensive trials on both the original and enriched versions of the VQA-Med CLEF 2019 dataset. Performance was assessed for VQA, response generation, and general diagnostic reasoning under both single-turn and multi-turn consultation scenarios. Comparative analyses against baseline and prompt-based methods were performed using standard metrics including accuracy, precision, recall, and F1-score.

To ensure comprehensive evaluation:

- Recall gauges the model's capacity to retrieve all pertinent answers from the dataset
- Accuracy measures the proportion of predictions that exactly match the ground-truth answers
- Precision measures the proportion of pertinent predictions among all predictions made
- F1-score, the harmonic mean of precision and recall, provides a balanced measure of both.

These metrics were calculated for VQA tasks at the question level by comparing generated responses with the VQA-Med 2019 dataset reference annotations. Accuracy was calculated as exact-match scores for key clinical statements, while precision, recall, and F1-score were calculated based on the retrieval of medical entities and findings for diagnostic report generation. The formulas used were:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{I}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{II}$$

$$\text{F1} - \text{score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{III}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{IV}$$

Where *FP* = False positives, *FN* = False negatives, *TP* = True positives, and TN = True negatives.

This thorough analysis highlights the benefits of LoRA-MedSim in terms of accuracy and robustness, providing a fair comparison with baseline and prompt-based approaches. In particular, balancing recall and precision is crucial for reliable medical decision-making. These metrics were applied to both baseline and prompt-based approaches across single-turn and multi-turn consultation settings, ensuring a fair comparison of model reliability.

4.1. DATASETS

The VQA-Med CLEF 2019 dataset, consisting of 4,200 medical images and over 15,000 question–answer pairs spanning four categories of clinical questions—organ identification, modality recognition, plane orientation, and abnormality detection—served as the foundation for model training and evaluation. To better reflect real-world diagnostic scenarios, the dataset was enriched with simulated multi-turn patient queries and synthetic laboratory results (e.g., ALT, aspartate aminotransferase, white blood cell count). This augmentation enabled the model to reason across a broader variety of clinical inputs while retaining the original dataset structure.

Additionally, a distinct synthetic patient-query test set was introduced to assess conversational robustness, while the original validation and test splits from CLEF 2019 were preserved to maintain comparability.

4.2. IMPLEMENTATION DETAILS

The proposed model was refined using LoRA on top of the large-scale VLM Qwen2-VLM. LoRA adapters with a scaling

Table 1. Training progress at the early and final epochs

Epoch	Training loss	Validation loss	Validation accuracy	Validation F1-score
1	1.8	1.75	0.62	0.58
2	1.788	1.738	0.623	0.583
3	1.777	1.727	0.625	0.586
4	1.765	1.715	0.628	0.589
5	1.754	1.704	0.631	0.592
...	...	...	...	...
96	0.696	0.646	0.869	0.858
97	0.685	0.635	0.872	0.861
98	0.673	0.623	0.875	0.864
99	0.662	0.612	0.877	0.867
100	0.65	0.6	0.88	0.87

Table 2. Ablation study results

Configuration	Accuracy	Precision	Recall	F1-score
Full system (LoRA+multimodal+multi-turn)	0.88	0.89	0.88	0.87
Without pretraining on structured laboratory data	0.812	0.82	0.79	0.805
Without multi-turn dialogue (static question answering)	0.828	0.84	0.81	0.825

factor of α = 32 and a rank of r = 8 were inserted into the transformer decoder blocks' attention layers. Fine-tuning was performed with PyTorch using automated mixed-precision on a single NVIDIA A100 80GB GPU, with a batch size of eight and a learning rate of 5×10^{-5}, over 100 epochs.

Training dynamics were monitored by tracking F1-score, validation accuracy, training loss, and validation loss at each epoch. The model demonstrated consistent improvements across all criteria (Table 1 and Figure 2). Validation loss decreased from 1.75 to 0.60, while training loss dropped from 1.80 to 0.65. F1-score improved from 0.58 to 0.87, and validation accuracy increased from 62% to 88%, demonstrating strong convergence and generalization.

These results demonstrate the model's increasing ability to interpret multimodal medical inputs and provide clinically meaningful responses over time.

4.3. ABLATION STUDIES

The impact of individual components in the proposed system was evaluated through ablation studies focusing on two key areas: multimodal pretraining and multi-turn dialogue simulation. The objective was to quantify each component's contribution to overall performance and to isolate its effect on diagnostic reasoning (Figure 3).

First, the model was tested without pretraining on structured laboratory data, using only the original image–question pairs from the VQA-Med CLEF 2019 dataset. This configuration led to a considerable decline in performance across all metrics, suggesting that incorporating structured clinical variables such as laboratory values plays a vital role in improving diagnostic reasoning and the relevance of generated answers.

Second, a static one-turn question–answering configuration was tested by disabling the multi-turn discussion mechanism. Although this design remained ineffective, accuracy and F1-score decreased, especially for follow-up questions requiring contextual recall or symptom clarification.

By contrast, the complete system, which integrates multimodal input processing, multi-turn simulation, and LoRA-based fine-tuning, outperformed the reduced configurations (Table 2), thereby validating the design decisions of our framework.

4.4. COMPARISON WITH BASELINE METHODS

The performance of the proposed system was evaluated against several prior state-of-the-art models on the VQA-Med 2019 dataset, including MMBERT,[4] WSDAN,[6] and Med-PMC,[8] to assess its efficacy relative to existing approaches. These baseline models achieved intermediate accuracy and F1-score values ranging from 0.655 to 0.705 (Table 3).

Across all evaluation metrics, our LoRA-fine-tuned Qwen2-VLM model noticeably outperformed these baselines. In particular, it demonstrated superior multimodal understanding and contextual reasoning, achieving an F1-score of 0.87 and an accuracy of 88%. These improvements are largely attributable to our incorporation of multi-turn dialogue simulation, fine-grained image grounding, and structured laboratory data integration—features absent in the baseline models.

Table 3. Performance comparison with prior state-of-the-art models

Model	Accuracy	Precision	Recall	F1-score
MMBERT[1]	0.672	0.66	0.65	0.655
WSDAN (ITLTA)[2]	0.69	0.68	0.67	0.675
Med-PMC[3]	0.698	0.71	0.7	0.705
LoRA-Qwen2-VLM (Ours)	0.88	0.89	0.88	0.87

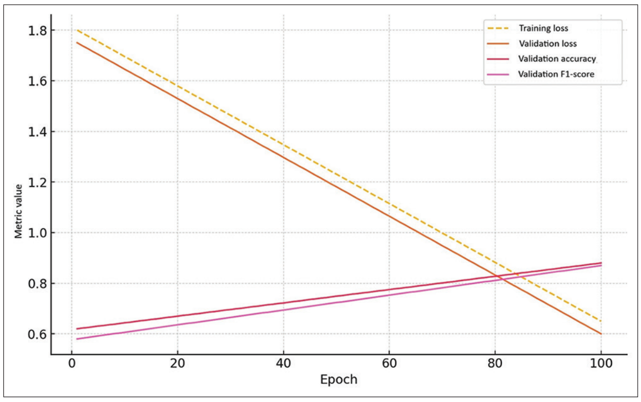

Figure 2. Training and validation trends over 100 epochs

The comparative evaluation results shown in Table 3 underscore the effectiveness of each design element within our system.

4.5. QUALITATIVE ANALYSIS

In addition to quantitative evaluation, a qualitative analysis was conducted to assess the linguistic clarity and clinical validity of the model's responses. Outputs from our model and a baseline prompt-based VLM were compared using the same multimodal clinical inputs (Table 4). The inputs consisted of a medical image (e.g., chest X-ray), laboratory results (e.g., ALT = 65.2 U/L), and a patient-initiated query such as: "What do my X-ray and lab report show?"

The baseline model generated a generic and ambiguous response: "Your results might be abnormal. Please speak with your doctor."

By contrast, our LoRA-fine-tuned Qwen2-VLM produced a thorough and clinically grounded response: "The X-ray shows mild cardiomegaly. With an increased ALT level of 65.2 U/L, hepatic stress may be the cause."

These findings demonstrate the model's capacity to combine data from multiple modalities and produce contextually relevant, educational responses, especially in

Table 4. Qualitative comparison of model responses

Input (image+lab+question)	Prompt-based Model	LoRA-Qwen2-VLM (Our model)
Chest X-ray+ALT=65.2 Q: What does my X-ray and lab report show?	Your results might be abnormal. Please consult your physician.	There is mild cardiomegaly visible in the X-ray. The ALT level is elevated at 65.2 U/L, which may indicate hepatic stress.
Abdominal CT+WBC=15.4 Q: Do I have signs of infection?	Further tests may be required to determine infection.	CT shows thickened intestinal walls, and WBC is elevated at 15.4 — findings suggest active infection.
Brain MRI+CRP=22 Q: Is there any inflammation or abnormality?	There might be something wrong, further review is needed.	MRI reveals white matter lesions, and CRP level of 22 mg/L supports inflammatory process, possibly early demyelination.

Abbreviations: ALT: Alanine aminotransferase; CRP: C-reactive protein; CT: Computed tomography; MRI: Magnetic resonance imaging; WBC: White blood cell (count).

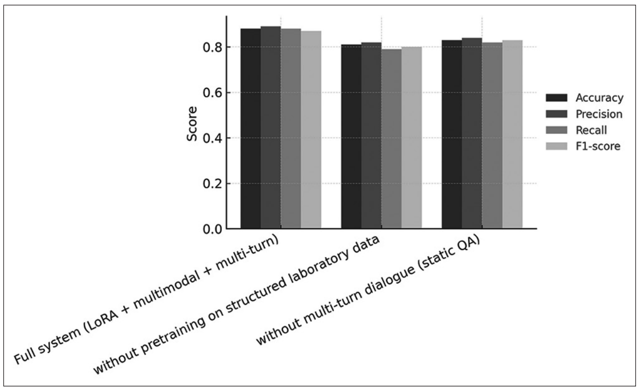

Figure 3. Performance metrics under ablation settings
Abbreviation: QA: Question answering

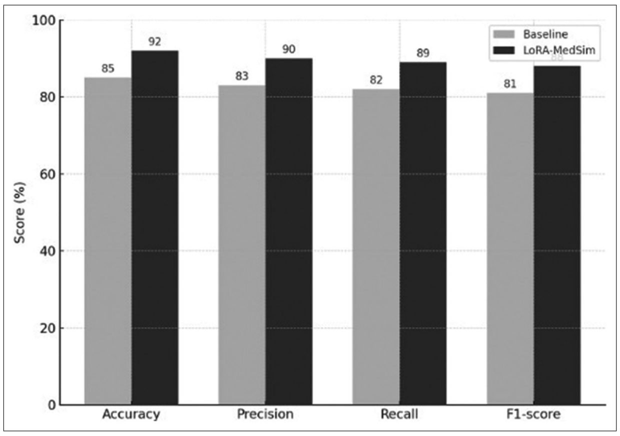

Figure 4. Performance comparison of baseline and LoRA-MedSim in multi-turn consultations

multi-turn scenarios where prior answers need to be recalled and updated.

By integrating heterogeneous inputs—including lab reports, radiological images, and patient histories—our model produced contextually relevant and instructive replies (Figure 4).

In multi-turn consultation situations, LoRA-MedSim achieved consistent improvements over prompt-based baselines: 7.8% in accuracy, 6.5% in precision, 6.1% in recall, and 6.3% in F1-score. Notably, the model correctly referenced a previously identified anomaly in 94% of cases when a follow-up query mentioned it, compared to 78% for the best-performing baseline.

These enhancements underscore the system's capacity to manage cross-modal reasoning while maintaining high diagnostic accuracy, memory of critical information, and conversational coherence—all of which are essential for realistic doctor–patient simulations.

5. CONCLUSION

The current study introduced LoRA-MedSim, a revolutionary multimodal architecture that integrates structured laboratory data, conversational reasoning, and radiological image interpretation within a single pipeline. A LoRA-fine-tuned Qwen2-VLM model was employed to simulate realistic doctor–patient interactions. On the VQA-Med CLEF 2019 benchmark, our approach achieved significant performance improvements. In multi-turn consultation scenarios, LoRA-MedSim outperformed baseline prompt-based models, with accuracy gains of up to 7.8% and F1-score improvements of 6.3%. These results underscore the model's capacity to retain and revise previous responses, thereby guaranteeing conversational coherence, which is essential for realistic and clinically applicable dialogue systems.

LoRA-MedSim advances prior work by integrating LoRA with multimodal reasoning in a way that enables scalable deployment, robust cross-modal integration, and efficient fine-tuning. Unlike earlier approaches that relied on static prompts or lacked multi-turn adaptability, our framework supports dynamic, context-aware interaction that more closely reflects real diagnostic workflows, addressing critical limitations in existing vision–language medical models.

Beyond gains in accuracy, LoRA-MedSim offers broader advantages in terms of adaptability across domains, flexibility, and potential for remote deployment in low-resource or home-based environments. Future research will concentrate on extending the system to cover a wider range of medical modalities (e.g., pathology slides, electrocardiograms), enhancing interpretability, and optimizing the framework for integration into real-time hospital workflows.

6. LIMITATIONS AND FUTURE WORK

Despite LoRA-MedSim's encouraging outcomes, a number of drawbacks must be noted. In this section, we highlight key issues related to deployment, interpretability, and generalizability, and outline possible avenues for further development.

6.1. GENERALIZATION ACROSS MODALITIES AND DEMOGRAPHICS

LoRA-MedSim has not yet been evaluated on non-English clinical data or datasets with broader demographic representation, despite its strong performance on VQA-Med 2019. The dataset comprises diverse radiology modalities—such as CT, MRI, X-ray, ultrasound, positron emission tomography, angiography, and mammography—but all questions and annotations are in English and derived from Western patient populations. This restricts the model's generalizability to underrepresented or non-English-speaking groups. To establish global clinical applicability, future research will focus on evaluating LoRA-MedSim with multilingual datasets and more demographically diverse populations.

6.2. INTERPRETABILITY AND EXPLAINABILITY

Although LoRA-MedSim demonstrates high accuracy in diagnostic reporting and clinical inquiry answering, interpretability is still a crucial barrier to clinical adoption. At present, the model functions largely as a black box, offering little transparency about how multimodal reasoning is performed.

Future research will investigate the integration of explainability techniques, such as Grad-CAM for visual attention over radiological images and attention weight visualization in the language decoder, to trace the model's decision-making processes. These methods could improve clinician trust and contribute to safer AI-assisted diagnosis.

6.3. DEPLOYMENT CHALLENGES

Although LoRA-MedSim was designed with adaptability in mind, several challenges complicate real-world deployment. The current system requires substantial GPU memory and may introduce delays in time-sensitive hospital workflows. Furthermore, handling sensitive patient data raises major privacy and regulatory compliance concerns.

Nonetheless, LoRA-MedSim has strong potential for patient-side deployment. The framework is particularly suited for home-based consultations, offering underserved or rural populations preliminary diagnostic support when doctors are not immediately available. This potential use case underscores the societal value of the system while mitigating several real-time clinical constraints.

Future research will investigate privacy-preserving approaches such as federated learning, along with lightweight deployment strategies including model quantization, knowledge distillation, and edge-device optimization.

ACKNOWLEDGMENTS

The authors are thankful to the National Telecommunication Institute in Cairo, Egypt, for the guidance and assistance offered in the laboratories.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHORS CONTRIBUTIONS

Conceptualization: Khadeja Fahmy, Mohamed Zorkany
Data curation: Khadeja Fahmy
Methodology: Mohamed Zorkany, Khadeja Fahmy
Supervision: Abd El-Hady Ammar
Writing–original draft: Khadeja Fahmy
Writing–review & editing: Mohamed Zorkany

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY STATEMENT

No new data were generated or analyzed in this study, and therefore data sharing is not applicable.

Submitted: 16 June 2025; Revision received: 11 August 2025; Accepted: 18 August 2025; Published: October 3, 2025

REFERENCES

1. Esteva A, Chou K, Yeung S, *et al*. Deep learning-enabled medical computer vision. *NPJ Digit Med*. 2021;4(1):5. doi: 10.1038/s41746-020-00376-2

2. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388(13):1201-1208. doi: 10.1056/NEJMra2302038

3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. doi: 10.1038/s41591-021-01614-0

4. Wang P, Zhang H, Yuan Y. Mcpl: Multimodal collaborative prompt learning for medical vision-language model. *IEEE Trans Med Imaging*. 2024;43(12):4224-4235. doi: 10.1109/TMI.2024.3418408

5. Li SS, Balachandran V, Feng S, *et al*. *Mediq: Question-Asking LLMs for Adaptive and Reliable Medical Reasoning*. [arXiv Preprint]; 2024.

6. Liu H, Liao Y, Ou S, *et al*. *Med-PMC: Medical Personalized Multi-Modal Consultation with a Proactive Ask-First-Observe-Next Paradigm*. [arXiv Preprint]; 2024.

7. Tiwari A. *Labour Monitoring in Pregnant women using Phonocardiography, Electrocardiography and Electromyography Technique*. [arXiv Preprint]; 2023.

8. Singhal K, Tu T, Gottweis J, *et al*. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31:943-950. doi: 10.1038/s41591-024-03423-7

9. Hu EJ, Shen Y, Wallis P, *et al*. *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations ICLR. Vol. 1. 2022. p. 3.

10. Kunisky D. Spectral pseudorandomness and the road to improved clique number bounds for Paley graphs. *Exp Math*. 2024:1-28. doi: 10.1080/10586458.2024.2400182

11. Khare Y, Bagal V, Mathew M, *et al*. MmBERT: Multimodal BERT pretraining for improved medical VQA. In: *Proceedings 2021 IEEE 18th International Symposium Biomed Imaging (ISBI)*. United States: IEEE; 2021. p. 1033-1036.

12. Abacha AB, Hasan SA, Datla VV, Liu J, Demner-Fushman D, Muller H. *VQA-Med: Overview of the Medical Visual Question Answering Task at Imageclef 2019*. Geneva: Zenodo.

13. Hu T, Qi H, Huang Q, Lu Y. *See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification*. [arXiv Preprint]; 2019.

14. ImageCLEF. *VQA-Med 2019 Dataset*. Available from: https://zenodo.org/records/10499039 [Last accessed on 2025 Jul 30].

15. Gapyak V, Rentschler CE, März T, Weinmann A. Anℓ^1-plug-and-play approach for MPI using a zero shot denoiser with evaluation on the 3D open MPI dataset. *Phys Med Biol*. 2025;70(2):025028. doi: 10.1088/1361-6560/ada5a

16. Liu H, Liao Y, Ou S, Wang Y, Liu H, Wang Y. *Med-PMC: Medical Personalized Multi-Modal Consultation with a Proactive Ask-First-Observe-Next Paradigm*. [arXiv Preprint]; 2024.

17. Liu C, OuYang C, Cheng SB, Shah A, Bai W, Arcucci R. *G2D: From Global to Dense Radiography Representation Learning Via Vision-Language Pre-Training*. United States: Cornell University; 2023.

18. Qin J, Liu C, Cheng S, Guo Y, Arcucci R. Freeze the backbones: A parameter-efficient contrastive approach to robust medical vision-language pre-training. In: *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. United States: IEEE; 2024. p. 1686-1690.

19. Liu C, Cheng S, Shi M, Shah A, Bai W, Arcucci R. IMITATE: Clinical prior guided hierarchical vision-language pre-training. *IEEE Trans Med Imaging*. 2024;44:519-529. doi: 10.1109/TMI.2024.3449690

20. Bai J, Bai S, Chu Y, *et al*. *Qwen Technical Report*. [arXiv Preprint]; 2023.