Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Medical Visual Question Answering for Telemedicine: A Secondary Data Review and the TRACE-MedVQA Framework

Dr Kazi Abdul Mannan

drkaziabdulmannan@gmail.com

Center for Academic & Professional Career Development and Research (CAPCDR)
https://orcid.org/0000-0002-7123-132X

Sidratul Muntaha Upoma

Shanto-Mariam University of Creative Technology https://orcid.org/0009-0006-8634-5043

Mohammad Arman Hossain

Shanto-Mariam University of Creative Technology

Research Article

Keywords: Medical visual question answering, telemedicine, multimodal learning, clinical decision support, explainable AI, retrieval-augmented generation

Posted Date: September 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10907584/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Medical visual question answering (Med-VQA) combines medical image understanding and natural-language processing to answer clinically relevant questions about radiographs, computed tomography, magnetic resonance imaging, pathology, and other visual data. This study evaluates existing Med-VQA methods from secondary sources and proposes a telemedicine-oriented framework. A qualitative comparative review was conducted across benchmark datasets, task-specific convolutional and transformer models, domain-adapted vision-language encoders, and recent generative large multimodal models. The evidence shows a clear progression from closed-set classification toward open-ended generation, but clinical deployment remains restricted by small and imbalanced datasets, weak external validation, hallucination risk, limited uncertainty reporting, and insufficient visual grounding. To address these limitations, the study proposes TRACE-MedVQA, a Telemedicine-Ready, Retrieval-Augmented, Calibrated, and Explainable framework. It combines modality-aware image encoding, clinical question encoding, cross-modal attention, trusted knowledge retrieval, dual answer routing, confidence calibration, abstention, visual explanation, and clinician oversight. The framework offers a practical research direction for safer remote decision support while preserving human responsibility.

1. Introduction

Telemedicine increasingly depends on digital images, electronic records, and remote communication to extend clinical expertise beyond the physical hospital. The World Health Organization places telemedicine, artificial intelligence, health information systems, and interoperable data within the wider digital-health ecosystem and emphasizes that these technologies should improve access, quality, and continuity of care [1]. Within this setting, Medical Visual Question Answering (Med-VQA) is an emerging multimodal task in which a system receives a medical image and a natural-language question and produces a clinically meaningful answer.

Unlike ordinary image classification, Med-VQA can respond to specific requests such as whether an abnormality is present, which organ is shown, where a lesion is located, or what imaging modality was used. The task therefore connects computer vision, natural-language processing, medical knowledge, and interactive decision support. Early benchmark datasets, including VQA-RAD, established clinically generated image-question-answer pairs [2], while PathVQA expanded the task to pathology images [3] and SLAKE added bilingual questions, semantic labels, multiple modalities, and a structured knowledge base [4].

Despite rapid technical progress, current Med-VQA systems are not yet reliable enough for routine telemedicine. Dataset scale remains small compared with general-domain vision-language corpora, and many benchmarks contain repeated answer patterns, class imbalance, limited demographic representation, or questions produced from captions rather than real consultations [2]-[5]. Classification-based systems can perform well on common yes/no or short-answer questions but cannot naturally

handle unseen answers. Generative systems provide flexible responses but may produce fluent statements that are unsupported by the image.

Published scores are also difficult to compare because studies use different train-test splits, preprocessing rules, answer vocabularies, and evaluation metrics. Patient-level leakage, inadequate external testing, and reporting only overall accuracy can hide clinically important failure modes. In telemedicine, these problems are amplified because images may be compressed, captured under variable conditions, transmitted through limited bandwidth, or reviewed without immediate access to a specialist. A wrong but confident answer could delay referral, create false reassurance, or increase unnecessary intervention. Therefore, performance alone is insufficient; a deployable system must also measure uncertainty, explain visual evidence, protect privacy, and defer difficult cases to clinicians.

A secondary-data evaluation is appropriate because the field already contains diverse datasets and architectures but lacks a unified interpretation focused on remote care. Existing surveys describe datasets and model families [5], whereas this study reorganizes the evidence around telemedicine requirements: robustness, efficiency, open-ended communication, traceability, calibration, and human oversight. The proposed framework is not intended to replace diagnosis. Instead, it defines how proven components can be integrated into a safer research prototype that supports clinicians during image-based remote consultation. The review is also timely because multimodal systems are moving from experimental answer classification toward conversational assistants, while governance guidance increasingly requires stronger evidence and risk controls. A structured synthesis can prevent telemedicine research from equating fluent output with clinical reliability.

The primary objective is to evaluate existing Med-VQA methods using published secondary evidence. Specific objectives are to compare datasets, visual and textual encoders, fusion strategies, output modes, and reported evaluation practices; identify methodological and deployment limitations; determine which design principles are most relevant to telemedicine; and propose an integrated framework that supports both closed-ended and open-ended questions while providing retrieval evidence, confidence estimates, visual grounding, and an abstention mechanism.

This study contributes a deployment-oriented synthesis rather than another isolated model comparison. It links technical choices to practical telemedicine risks and proposes a reproducible evaluation plan. For researchers, the framework identifies variables and metrics that should be reported beyond accuracy. For clinicians and institutions, it clarifies why human review, audit logs, privacy controls, and calibrated uncertainty are necessary. For low-resource settings, it emphasizes efficient encoders and selective generation so that useful assistance can be developed without assuming continuous access to large computing infrastructure.

2. Literature Review

Med-VQA developed from general visual question answering but faces substantially different constraints. Medical images are specialized, abnormalities may occupy small regions, clinical

terminology is precise, and plausible-looking answers can still be unsafe. VQA-RAD contains 315 radiology images and 3,515 clinician-generated questions and answers, making it highly valuable for clinical relevance but small for deep learning [2]. PathVQA contains approximately 32,799 questions from 4,998 pathology images, created through a semi-automated pipeline and manual verification [3]. SLAKE contains 642 images and 14,028 bilingual question-answer pairs across radiography, CT, and MRI, with semantic labels and a medical knowledge base [4].

ImageCLEF VQA-Med further encouraged comparable competition-based evaluation; in 2021, the best submission reached 0.382 accuracy and 0.416 BLEU on abnormality-focused tasks, illustrating the difficulty of specialized open answers [6]. The later PMC-VQA resource greatly increased scale to about 227,000 question-answer pairs from 149,000 images and enabled instruction-tuned generative research [7]. These datasets collectively support progress, but they vary in source, annotation process, modality, question distribution, and clinical realism. Question composition also matters.

Benchmarks commonly separate closed-ended questions, such as yes/no or categorical recognition, from open-ended questions that require a term, phrase, or explanation. Closed questions are easier to score but can encourage language shortcuts, whereas open questions better test clinical vocabulary yet introduce synonym and phrasing problems. Consequently, exact-match accuracy can underestimate a semantically correct answer, while BLEU can overvalue surface overlap. Dataset design therefore influences not only training but also what researchers believe the model has learned.

The first major approach treated Med-VQA as answer classification. A convolutional neural network extracted image features, an LSTM or recurrent encoder represented the question, and a fusion module predicted one answer from a fixed vocabulary. ImageCLEF 2021 systems illustrate this design. The TAM model combined a modified ResNet-34, LSTM question features, multimodal factorized bilinear pooling, and co-attention [8]. The Yunnan University system used visual features, BioBERT-based language representation, multi-modal factorized high-order pooling, co-attention, and answer classification [9]. Such systems are computationally manageable and often strong for frequent closed-ended answers. However, their output space cannot naturally produce unseen clinical phrases, and performance may be dominated by common answer classes.

A second line of research improved representation learning and cross-modal interaction. PubMedCLIP adapted CLIP to biomedical image-text pairs derived from PubMed-related resources and improved overall Med-VQA accuracy by up to three percentage points over previously used visual encoders in tested settings [10]. This finding supports domain-specific pretraining: visual features learned jointly with medical language are more useful than generic visual features alone. Transformer-based fusion further improved the interaction among image regions, question tokens, and candidate answers.

Q2ATransformer introduced learnable answer embeddings and an answer-querying decoder that preserved classification efficiency while allowing richer interaction with open-answer information. It reported 79.19% open-ended accuracy on VQA-RAD and 54.85% on PathVQA, with large gains over earlier baselines [11]. LaPA later used latent prompts and prior disease-organ knowledge to filter

clinically relevant information, reporting improvements on VQA-RAD, SLAKE, and VQA-Med 2019 [12]. These models show that attention, domain knowledge, and answer-aware reasoning can reduce the limitations of simple feature concatenation.

The third stage reframed Med-VQA as language generation. Prefix-tuning methods mapped visual features into learnable language-model tokens, enabling parameter-efficient open-ended answering on small domain datasets [13]. PMC-VQA and MedVInT aligned a pretrained vision encoder with a large language model and used large-scale visual instruction data to improve free-form answers across VQA-RAD, SLAKE, and ImageCLEF benchmarks [7]. LLaVA-Med followed a biomedical curriculum: it first aligned biomedical vocabulary using figure-caption pairs and then learned instruction-following conversation, demonstrating that general multimodal models can be adapted to biomedical images with targeted data [14]. Generative systems are better suited to natural teleconsultation because users are not restricted to a fixed answer vocabulary. Nevertheless, automatic text metrics may reward lexical similarity without establishing clinical correctness. These systems can hallucinate findings, ignore subtle image evidence, or use memorized language priors when the image is ambiguous.

Recent research therefore emphasizes robustness, explainability, and retrieval. MedThink was designed to improve the transparency of medical visual reasoning rather than return only an answer [15]. Noise-aware representation learning introduces denoising before visual-prefix mapping and parameter-efficient adaptation, directly addressing sensitivity to irrelevant visual changes [16]. Retrieval-augmented approaches can connect the image-question pair to verified external evidence instead of relying only on model parameters. This direction is especially important for telemedicine, where clinicians need an answer that can be checked against a trusted source. Yet retrieval does not automatically guarantee safety: irrelevant evidence, outdated guidelines, or weak image-grounding can still produce a persuasive error.

Across these stages, five analytical dimensions are most important. First, dataset validity determines whether reported performance transfers to real remote consultation. Second, encoder quality determines whether subtle abnormalities and medical terms are represented. Third, fusion determines whether the answer actually depends on both the image and question. Fourth, output design creates a trade-off between closed-set reliability and generative flexibility. Fifth, evaluation determines whether the system is clinically trustworthy. The literature strongly evaluates accuracy and language similarity but less consistently reports calibration, subgroup performance, out-of-distribution robustness, visual grounding, latency, or clinically severe errors. Consequently, a model may be technically competitive without being telemedicine-ready.

The central research gap is not the absence of another powerful encoder; it is the absence of an integrated, deployment-oriented design that combines clinical retrieval, calibrated confidence, abstention, visual evidence, efficient execution, and clinician control. Many task-specific models are accurate but closed-set, while large multimodal models are conversational but difficult to verify. Benchmark studies rarely evaluate compressed images, external hospitals, patient-level splits,

demographic subgroups, or referral decisions. A framework is therefore needed that routes different question types to different answer mechanisms and refuses unsafe predictions when evidence is inadequate.

The literature demonstrates a consistent evolution from CNN-RNN classification to attention-based fusion, biomedical contrastive pretraining, answer-aware transformers, and instruction-tuned generative models. Each stage solves part of the problem, but no single approach fully satisfies remote-care requirements. Classification offers control, generation offers flexibility, retrieval offers evidence, and calibration offers risk awareness. Their combination, supported by visual explanation and human oversight, provides the strongest conceptual basis for a telemedicine-oriented Med-VQA system.

3. Theoretical and Conceptual Framework

The theoretical framework is based on multimodal representation learning, evidence-based decision support, and human-in-the-loop control. Multimodal representation learning assumes that a correct answer depends on the joint relationship among visual features, question meaning, and relevant medical knowledge rather than on any modality independently. Let an image be I and a question be Q. A visual encoder produces v = E_v(I), while a clinical language encoder produces q = E_q(Q). A cross-modal function F combines them as h = F(v, q), allowing question tokens to attend to diagnostically relevant visual regions. This reflects the attention-based progression reported in task-specific Med-VQA research [8], [11].

Evidence-based decision support adds a retrieval function R that selects trusted knowledge k from an approved repository using the joint representation h. The system therefore estimates an answer from image evidence, question intent, and retrieved clinical context rather than relying solely on memorized parameters. Human-in-the-loop theory defines the model as an assistive component: uncertain, contradictory, or high-risk cases are escalated to a qualified clinician. This is consistent with WHO guidance that health AI should protect autonomy, promote safety, remain transparent, and preserve accountability [17]. Together, these principles establish that a telemedicine Med-VQA answer should be accurate, evidence-linked, calibrated, and reviewable.

The proposed conceptual framework is named TRACE-MedVQA: Telemedicine-Ready, Retrieval-Augmented, Calibrated, and Explainable Medical Visual Question Answering. It has four connected layers. The input-governance layer performs de-identification, image-quality checking, modality detection, and question-type identification. The multimodal intelligence layer uses a domain-adapted vision encoder, a clinical text encoder, and cross-modal attention. The evidence-and-answer layer retrieves approved knowledge and routes closed-ended questions to a constrained classifier while routing open-ended questions to a controlled generator. The safety-and-delivery layer calibrates confidence, checks visual grounding and answer consistency, abstains when risk exceeds a threshold, displays supporting regions and sources, and records clinician feedback. The integrated workflow is shown in Fig. 1.

3.1 Mathematical Formulation

For a closed-ended question, the answer distribution is p_c = softmax(W_c[h; k] + b_c). For an open-ended question, the generator estimates P(y_t | y_<t, h, k) while constrained decoding discourages unsupported terms. Confidence is calibrated on validation data, and the system abstains when max(p_c) < τ, retrieval relevance is low, image quality is inadequate, or generation-consistency checks fail. A combined training objective can be written as L = λ1L_cls + λ2L_gen + λ3L_align + λ4L_cal + λ5L_ground, where the terms represent classification, generation, image-text alignment, calibration, and visual grounding. This formulation allows performance and safety to be optimized jointly.

4. Methodology

This study uses a qualitative secondary-data design. It does not collect patient data or claim new clinical performance. The unit of analysis is a published Med-VQA dataset, model, benchmark report, or authoritative digital-health guidance document. Sources were selected when they described a recognized dataset, introduced an influential architecture, reported comparative results, or addressed governance relevant to medical multimodal systems. The review covers the progression from early radiology and pathology datasets to competition models, biomedical contrastive learning, transformer fusion, open-ended generation, and large multimodal instruction tuning.

A structured extraction matrix was designed with the following fields: publication year; dataset and modality; question type; data scale; image encoder; language encoder; fusion or attention method; output mode; use of external knowledge; reported metric; explainability; uncertainty handling; computational demand; and stated limitation. The analytical process had three stages. First, studies were grouped by output paradigm: fixed-answer classification, hybrid answer-aware modeling, and generative answering. Second, their reported strengths and weaknesses were compared across the extraction fields. Third, the evidence was mapped to telemedicine requirements, including variable image quality, limited computing resources, need for natural interaction, auditability, privacy, and timely referral.

The proposed TRACE-MedVQA workflow was derived through design synthesis. Components were included only when the reviewed literature identified a repeated limitation or a validated technical direction. Domain-adapted vision-language encoders address representation mismatch [10]. Cross-modal attention addresses weak interaction between image and question [8], [11]. Retrieval addresses the need for current, checkable evidence. Dual answer routing preserves the reliability of classification for constrained questions while retaining generation for explanations. Calibration and abstention address overconfidence, while grounding and clinician review address traceability.

This method produces a conceptual framework and evaluation protocol rather than an experimentally trained model. To reduce selection bias, the synthesis gives priority to original dataset papers, benchmark overviews, model-introduction papers, and official governance documents. Reported numerical results are interpreted only within their own experimental settings; no pooled meta-analytic

effect is calculated because datasets, splits, answer normalization, and metrics are heterogeneous. The comparison therefore focuses on patterns that recur across studies rather than declaring one universal best model. A study is considered more deployment-relevant when it reports clinically sourced questions, external or patient-level validation, error categories, uncertainty, grounding, efficiency, or human evaluation.

4.1 Variable Selection

The main independent design variables are visual encoder type, text encoder type, fusion mechanism, retrieval availability, answer mode, calibration method, and explanation mechanism. Context variables include dataset size, modality, question category, image quality, answer imbalance, and computational setting. Dependent outcomes include closed-ended accuracy, open-ended clinical correctness, macro F1-score, calibration error, hallucination rate, grounding quality, latency, and abstention coverage. Deployment-readiness indicators include external validation, patient-level separation, privacy protection, audit logs, and human-review support.

4.2 Cross Validation

Although the present study is qualitative, the proposed framework should be tested using patient-level stratified five-fold cross-validation. All images and questions from one patient or source study must remain in a single fold to prevent information leakage. Four folds are used for training and internal validation and one fold for testing, rotating until every fold has served as the test set. Hyperparameters and confidence thresholds must be selected only inside the training folds. A final untouched external dataset from another institution, device, or geographic setting should then assess generalizability. Results should be reported as mean, standard deviation, and confidence intervals, with separate analyses for modality, question type, and subgroup.

4.3 Model Evaluation

Closed-ended answers should be evaluated with accuracy, macro precision, recall, F1-score, and confusion matrices. Open-ended answers require BLEU or ROUGE for comparability, but these should be supplemented with BERTScore, clinically normalized exact match, and blinded expert ratings of correctness, relevance, and harmfulness. Calibration should be measured with expected calibration error and reliability diagrams. Selective prediction should report risk-coverage curves showing performance as uncertain cases are rejected. Visual explanations should be evaluated for localization agreement when annotations exist and by clinician review otherwise. Robustness tests should include compression, noise, missing context, paraphrased questions, and out-of-distribution images. Latency and memory use are necessary for telemedicine feasibility.

4.4 Software

A future implementation can use Python, PyTorch, Hugging Face Transformers, OpenCLIP, scikit-learn, and Captum or Grad-CAM for interpretation. FAISS can support retrieval from a curated knowledge base.

Experiments should use version-controlled code, fixed random seeds, configuration files, and documented hardware. Microsoft Word and a reference manager can be used for article preparation and IEEE citation management.

4.5 Ethical Issue

Only de-identified, lawfully accessible data should be used. Data licenses, institutional approvals, and secondary-use restrictions must be respected. The model should not provide autonomous diagnosis or treatment. WHO guidance for health AI and large multimodal models emphasizes safety, transparency, accountability, inclusiveness, privacy, and meaningful human control [17]. Development should include bias testing, cybersecurity, clinician oversight, user warnings, incident reporting, and protection against storing identifiable images or conversations. Good machine-learning practice also requires representative data, independent test sets, clear human-AI team performance, and monitoring after deployment [18].

5. Results

The secondary-data synthesis produced four principal findings. First, dataset growth has improved model development but has not eliminated the validity problem. VQA-RAD offers clinician-authored questions but is extremely small [2]. PathVQA provides more examples but is derived largely from educational sources [3]. SLAKE adds bilingual and knowledge-enhanced annotations yet still contains only hundreds of images [4]. PMC-VQA increases scale dramatically through automated construction [7], but scale does not guarantee that questions reflect real telemedicine dialogue. Therefore, dataset quantity and clinical realism must be evaluated separately.

Second, domain adaptation and stronger fusion consistently improve task-specific systems. ImageCLEF approaches demonstrated that residual visual encoders, biomedical language encoders, bilinear pooling, and co-attention can extract useful cross-modal patterns [8], [9]. PubMedCLIP showed measurable benefit from medical image-text pretraining [10]. Q2ATransformer and LaPA further demonstrated that answer-aware attention, latent prompts, and prior knowledge can improve difficult open-ended categories [11], [12]. The evidence supports retaining specialized, efficient components rather than replacing all task-specific models with a single unrestricted generator.

Third, generative instruction-tuned systems improve natural interaction and answer flexibility. Prefix tuning, MedVInT, and LLaVA-Med reduce dependence on fixed answer vocabularies and make multi-turn or explanatory use more feasible [7], [13], [14]. However, the reviewed evidence does not establish that fluent generation is consistently clinically safe. Standard language metrics do not directly measure unsupported findings, referral consequences, or the severity of errors.

Fourth, deployment-oriented measures are underreported. Few studies jointly evaluate calibration, abstention, external validity, visual grounding, latency, subgroup fairness, and image degradation. These gaps justify TRACE-MedVQA. Its dual-router design assigns predictable closed questions to

classification and flexible questions to constrained generation. Retrieval, calibration, quality gates, visual evidence, and clinician escalation are added because no reviewed model family alone addresses all these requirements. A fifth finding concerns evaluation mismatch. Strong results on one benchmark do not imply stable performance across another because answer distributions, imaging modalities, and annotation styles differ.

Improvements from biomedical pretraining are important, but the literature also shows that the same visual backbone can behave differently across datasets [10]. Similarly, a model that generates a medically related phrase may receive a reasonable language score even when the phrase is too broad, misses laterality, or invents severity. For telemedicine, evaluation must distinguish minor wording variation from clinically consequential error. The synthesis therefore supports reporting category-level outcomes, confidence calibration, abstention performance, and expert-rated severity in addition to aggregate scores.

6. Discussion

The findings indicate that telemedicine changes the definition of a successful Med-VQA system. On a benchmark, success may mean the highest aggregate accuracy or BLEU score. In remote care, success means giving a useful answer when evidence is adequate, refusing when evidence is insufficient, and helping the clinician understand why. This distinction explains why older classification models remain valuable. Their vocabulary is limited, but their outputs can be controlled and calibrated. A yes/no abnormality question, modality question, or anatomical category may be safer through a constrained head than through unrestricted language generation.

Generative models are nevertheless necessary for questions that request descriptions, comparisons, or explanations. The proposed dual router therefore avoids treating classification and generation as competing paradigms. It uses each where its risk-benefit profile is strongest. Retrieval also changes the system from a memory-only predictor into an evidence-supported assistant. In practice, retrieved material must come from approved and versioned sources, and the displayed answer should distinguish image-derived observations from general medical information.

The proposed framework has practical relevance for low-resource telemedicine. Lightweight encoders, parameter-efficient fine-tuning, compressed retrieval indexes, and selective generation can reduce hardware demands. The model can prioritize triage support, image-quality feedback, anatomical orientation, common abnormality questions, and referral prompts rather than attempt unrestricted diagnosis. The abstention threshold can be adjusted according to clinical risk: a low-risk educational question may tolerate broader generation, while a possible emergency finding should require strong evidence and clinician confirmation.

This study has limitations. It is a narrative comparative synthesis rather than a formal meta-analysis, and published metrics are not directly comparable across datasets and splits. Recent large multimodal models evolve rapidly, and some studies rely on preprints. The proposed framework has not been trained

or prospectively tested. Consequently, TRACE-MedVQA should be interpreted as a research design derived from recurring evidence, not as proof of clinical effectiveness. Its value lies in defining measurable requirements for the next empirical stage.

Another implication is that the interface is part of model safety. A telemedicine user should not receive a single unqualified sentence. The interface should show whether the answer was classified or generated, whether image quality passed the gate, which visual region influenced the response, what evidence was retrieved, and whether the system recommends specialist review. When confidence is low, a useful response may be a clarification request or a statement that the image is insufficient rather than a guess. This design reduces the risk that conversational fluency is mistaken for certainty.

The framework also encourages feedback without silently retraining on every interaction. Clinician corrections should enter a reviewed audit dataset, be checked for privacy and label quality, and be used in controlled model updates. This supports monitoring for drift and recurring failure categories while preserving accountability. In this sense, TRACE-MedVQA is not only a neural architecture; it is a lifecycle that connects data governance, model evaluation, interface communication, and clinical supervision.

7. Conclusion

This article evaluated Med-VQA methods from secondary evidence and proposed TRACE-MedVQA, a telemedicine-ready framework integrating specialized multimodal learning, trusted retrieval, dual answer routing, calibration, explanation, abstention, and clinician oversight.

The objectives were achieved by comparing datasets, encoders, fusion strategies, output paradigms, and evaluation practices; identifying deployment gaps; and translating the findings into an integrated conceptual workflow and test protocol.

The review found that benchmark development progressed from small clinically valuable datasets to larger automatically constructed resources. CNN-RNN and attention-based classifiers remain efficient and controllable, while domain-specific pretraining and answer-aware transformers improve multimodal representation. Instruction-tuned large multimodal models enable flexible conversation but introduce verification and hallucination challenges. Across model families, uncertainty calibration, external validation, robustness, visual grounding, and clinically meaningful error analysis remain insufficiently reported. These findings show that no single architecture currently satisfies all telemedicine requirements.

The study contributes a deployment-centered analytical framework and a new conceptual model. TRACE-MedVQA combines components that are often studied separately: input quality control, modality-aware encoding, clinical question representation, cross-modal attention, curated retrieval, classification-generation routing, confidence calibration, visual explanation, and human escalation. It also defines variables, patient-level cross-validation, external testing, selective prediction, expert assessment, and operational metrics that can guide reproducible evaluation. The contribution is conceptual and

methodological, providing a structured basis for future implementation. By explicitly separating conceptual promise from measured clinical effectiveness, the framework also discourages exaggerated claims and makes the next validation steps visible.

Health institutions considering Med-VQA should require transparent intended-use statements, approved data governance, patient-level and external validation, subgroup reporting, calibrated uncertainty, and clear escalation pathways. Generated answers should never be presented as independent diagnoses. Policies should require users to see confidence, evidence provenance, and limitations, while audit logs record model versions, retrieved sources, outputs, and clinician actions. Procurement standards should also consider latency, cybersecurity, update management, and human-AI team performance rather than headline accuracy alone. These measures align technical development with responsible digital-health governance.

Future work should implement TRACE-MedVQA on VQA-RAD, SLAKE, PathVQA, and an institutionally approved telemedicine dataset. Research should compare classifier-only, generator-only, and dual-router systems under patient-level cross-validation and external testing. Additional experiments should measure robustness to compression, device differences, incomplete clinical context, adversarial prompts, and demographic shift. Clinicians should evaluate answer correctness, harmfulness, usefulness, and grounding. Multilingual adaptation, including Bangla questions, is particularly relevant for equitable remote care. Prospective studies should finally assess whether the framework improves consultation time, referral quality, and clinician confidence without increasing automation bias. Federated or privacy-preserving learning may support improvement without centralizing data, but it requires evaluation of communication cost and governance.

References

1. World Health Organization (2025) Global Strategy on Digital Health 2020–2027. WHO, Geneva, Switzerland
2. Lau JJ et al (2018) A dataset of clinically generated visual questions and answers about radiology images, Scientific Data, vol. 5, Art. no. 180251. 10.1038/sdata.2018.251
3. He X, Zhang Y, Mou L, Xing E, Xie P PathVQA: 30000 + questions for medical visual question answering, arXiv:2003.10286, 2020.
4. Liu B, Zhan L-M, Xu L, Ma L, Yang Y, Wu X-M (2021) SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in Proc. IEEE 18th Int. Symp. Biomedical Imaging (ISBI), 2021, pp. 1650–1654
5. Lin Z et al (2023) Medical visual question answering: A survey, Artificial Intelligence in Medicine, vol. 143, Art. no. 102611
6. Ben Abacha A, Sarrouti M, Demner-Fushman D, Hasan SA, Müller H (2021) Overview of the VQA-Med task at ImageCLEF 2021: Visual question answering and generation in the medical domain, CEUR Workshop Proceedings, vol. 2936, pp. 1081–1088

7. Zhang X et al (2023) PMC-VQA: Visual instruction tuning for medical visual question answering, arXiv:2305.10415
8. Li Y, Yang Z, Hao T (2021) TAM at VQA-Med 2021: A hybrid model with feature extraction and fusion for medical visual question answering, CEUR Workshop Proceedings, vol. 2936, pp. 1295–1304
9. Xiao Q, Zhou X, Xiao Y, Zhao K (2021) Yunnan University at VQA-Med 2021: Pretrained BioBERT for medical domain visual question answering, CEUR Workshop Proceedings, vol. 2936, pp. 1405–1411
10. Eslami S, Meinel C, de Melo G PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? in Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2023, pp. 1181–1193. 10.18653/v1/2023.findings-eacl.88
11. Liu Y, Wang Z, Xu D, Zhou L (2023) Q2ATransformer: Improving medical VQA via an answer querying decoder, in Information Processing in Medical Imaging (IPMI 2023), pp. 445–456
12. Gu T, Yang K, Liu D, Cai W LaPA: Latent prompt assist model for medical visual question answering, arXiv:2404.13039, 2024.
13. van Sonsbeek T, Derakhshani MM, Najdenkoska I, Snoek CGM, Worring M (2023) Open-ended medical visual question answering through prefix tuning of language models, in Medical Image Computing and Computer Assisted Intervention (MICCAI 2023), pp. 726–736
14. Li C et al (2023) LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day, in Advances in Neural Information Processing Systems, Datasets and Benchmarks Track
15. Gai X et al MedThink: Explaining medical visual question answering via multimodal decision-making rationale, arXiv:2404.12372, 2024.
16. Pratama IPA, Ofoghi B, Sajjanhar A, Gao S Noise-aware visual representation learning for medical visual question answering, arXiv:2606.05535, 2026.
17. World Health Organization (2024) Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. WHO, Geneva, Switzerland
18. U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare products Regulatory Agency, Good machine learning practice for medical device development: Guiding principles, (2021)

Figures

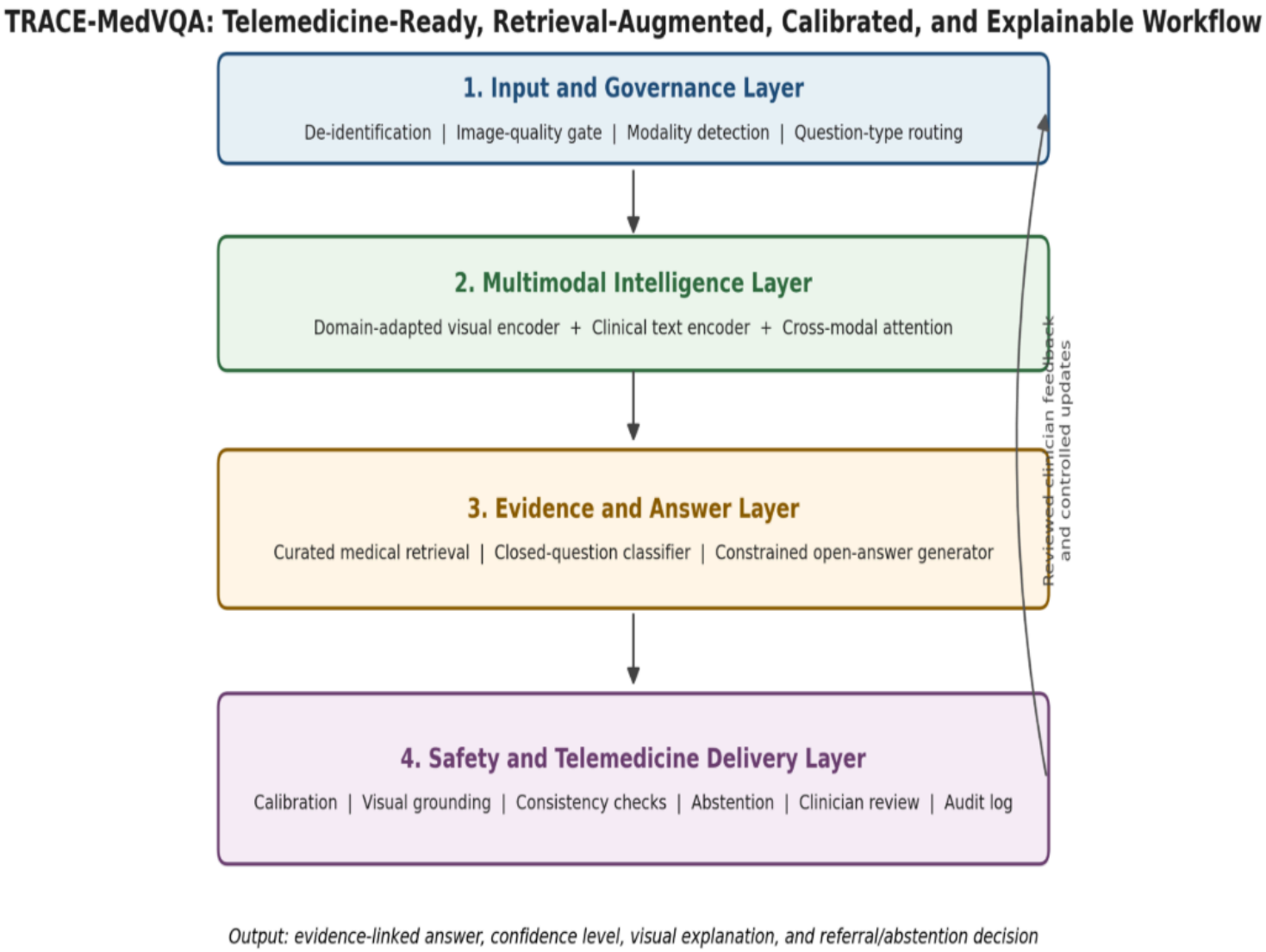


Figure 1

Integrated TRACE-MedVQA research framework and workflow.