Grenze International Journal of Engineering and Technology, June Issue

Enhancing Medical Visual Question and Answering through Domain-Specific LLMs & ILA Framework

Atla Hruthvika[1] and K. Radhika[2]
[1-2]M. Tech (AI & DS), Chaitanya Bharathi Institute of Technology, Hyderabad, India
Email: atlahruthvikareddy@gmail.com, Kradhika_aids@cbit.ac.in

***Abstract*— The Medical Visual Question Answering (MVQA) has become a research necessity in the creation of intelligent clinical decision-support systems, but the vast majority of current multimodal AI systems are black-box opaque models whose diagnostic results are fluent but cannot be factually verified. These ungrounded or hallucinated reactions are dangerous and unsafe in high risk health care environments. The Image-to-Label-to-Answer (ILA) framework proposed in this paper is a formalizable, auditable MVQA system which architecturally implements the decoupling between visual measurement and linguistic reasoning. It is based on the quantitative segmentation of Swin-UNETR and structured label generation and persistence, a two-stage domain-specific LLM pipeline (FLAN-T5 + BioMistral) and a safety-critical Factual Audit mechanism. The ILA framework, assessed on five multi-modal clinical datasets, obtains a mean DSC of 0.921 in segmenting organs, a 100 percent Factual Consistency Rate post-audit and a 96.4 percent end-to-end diagnostic accuracy across all modalities, indicating that structured evidence grounding can deliver accurate data, remove all hallucination as well as offer full diagnostic traceability.**

***Keywords*— Medical Visual Question Answering; Image-to-Label-to-Answer; Factual Grounding; Swin-UNETR; Domain-Specific LLMs; Hallucination Mitigation; Clinical Decision Support; Factual Audit.**

I. INTRODUCTION

Tomography (CT), Magnetic Resonance Imaging (MRI), plain X-rays and digital microscopy has become increasingly important in medical diagnosis. These modalities produce complex, high-dimensional data that must be accurately and consistently interpreted using a lot of clinical expertise. The large and increasing amount of medical imaging data and their intricacy has necessitated intelligent, automated decision-support systems. These systems help clinicians to save time on interpretation, minimize human error and enhance diagnostic consistency in different healthcare environments [1].

Medical Visual Question Answering (MVQA) has proved to be one of the most promising paradigms within this space. MVQA systems fill the gap between visual data complexity and actionable clinical information by allowing clinicians to ask free and natural language questions of medical images and get contextually relevant responses. Recent developments in multimodal deep learning, which pairs vision encoders with Large Language Models (LLMs), have resulted in systems that can generate very fluent and contextually plausible responses to queries on medical images [2].

Grenze ID: 01.GIJET.12.2.214_26

Nevertheless, a very severe and well-documented weakness of these systems is that they do not have explicit, traceable evidence grounding. State-of-the-art multimodal architectures combine image embeddings and language generation pipelines in a single end-to-end learned mapping. Although this makes it possible to speak fluent. Response generation, it does not give a way of checking the existence of the generated text being based on actual visual measurements. It is called hallucination and is a well-known safety issue in the implementation of medical AI [3].

Interpretability/ Verifiability are not design features in clinical settings, but a regulatory/ethical requirement. An AI diagnostic system, which comes up with certain answers that cannot be traced with supportive evidence, is not safe to be incorporated into clinical workflow. Regulatory authorities, medical professionals and hospital ethics committees demand that any AI-generated diagnostic results be subject to audit, challenge and trace to objective, reproducible measurements [4].

The fundamental limitation of current MVQA systems lies in lack of reasoning ability rather than lack of an intermediate representation between image-level evidence and text-level outputs that is organized and auditable. Even domain-specific LLMs fine-tuned on large biomedical corpora can make wrong clinical inferences unless based on structures that quantify visual observations. This fundamental gap is filled in the current paper by introducing Image-to-Label-to-Answer (ILA) framework - an innovative MVQA architecture that requires an opaque generation pipeline to be converted into a fully auditable, transparent, and evidence-based clinical decision-support system through the introduction of a quantitative intermediate labeling stage [1].The primary contributions of this work are:

- Explicit quantitative measurement which is imposed on a visual-linguistic architecture prior to any language generation.
- A visual-linguistic architecture which is decoupled and which requires explicit quantitative measurement prior to language generation.
- A mandatory intermediate structured labeling stage (C_{stor}^{ed}) that converts raw pixel measurements into immutable, persistent evidence prior to reasoning.
- Two step domain specific LLM pipe with FLAN-T5 to generate a draft, and BioMistral to refine the draft in a clinically accurate manner.
- A Factual Audit system that is safety critical, and explicitly cross-validates all generated statements with stored quantitative labels, pruning out hallucinated results.
- Scalable, cloud-deployable architecture that has been tested on CT, X-ray and microscopy modalities and is now prepared to be deployed in a clinical environment.

II. Related Works

The development of MVQA systems has gone through a series of generations of methodology. The first methods used Convolutional Neural Networks (CNNs) to extract visual features, and combine them with Long Short-Term Memory (LSTM) networks or text encoders to generate answers. Although they worked well on closed-domain classification tasks like VQA-Med 2018 and PathVQA, these systems provided limited reasoning capabilities and poor interpretability. Their responses were mostly conditioned based on features of global images without traces of space or locations [6].

Transformer-based architectures ushered in an enormous increase in MVQA capability. ViLBERT, LXMERT, and eventual clinical versions like ClinicalBERT-Visual are dual-encoder models that allow fine-grained prediction of image regions and text tokens by cross-modal attention. These models showed significantly better performance on open-ended medical VQA items. They are, however, black-box systems in nature - the reasoning processes that go on inside them are opaque and their outputs are not traceable to individual visual measures making them inappropriate to use in high-stakes clinical situations [7].

Swin-UNETR-type segmentation-oriented vision transformers and variants designed to address medical imaging have shown state-of-the-art results in 3D volumetric medical image segmentation. The models are also very good in delineating tumor boundaries, estimating organ volumes, and characterizing lesions regardless of the CT or MRI mode. Although they have good quantitative measurement abilities, previous studies have not directly incorporated such segmentation outputs as.grounding evidence for downstream language reasoning, representing a significant missed opportunity for factual grounding in MVQA [8].

The large-scale biomedical language models such as BioGPT, BioMistral, Med-PaLM, and Meditron have significantly contributed to the development of medical natural language understanding. Refined on clinical corpora (such as PubMed, MIMIC clinical notes, and medical QA datasets) the models have shown a very high performance on medical question answering benchmarks. Nevertheless, the problem of hallucination still exists:

in the absence of explicit visual grounding mechanisms, these language models produce plausible clinical language with no indication that it may be factual, related to the content of the image [9].
Hallucination has been partially solved with Retrieval-Augmented Generation (RAG) systems that augment the outputs of language models with or found external information in medical databases. Nevertheless, RAG systems recover textual knowledge instead of basing responses on quantitative visual representations based on the input image. Visualization and region-of-interest highlighting techniques enhance the interpretability of these techniques in a qualitative, non-quantitative fashion but not the layer of immutable evidence demanded by true clinical verifiability [10].
Three critical research gaps motivate the proposed framework:

- Current systems of MVQA make no stringent efforts to incorporate quantitative segmentation outputs as required evidence into their architectures.
- There is no existing framework that explicitly and automatically checks the statements of clinical statements with the image-based quantitative measurements.
- Lack of long-term and auditable intermediate representations implies that MVQA outputs cannot be tracked, disputed or confirmed by subsequent clinical assessment procedures.

The proposed ILA framework proposes a systematic architecture that imposes the quantitative measure, the label storage and the systematic verification of the facts as the mandatory pipeline steps instead of the optional post-factum additions, which the proposed architecture directly addresses the three gaps.

III. METHODOLOGY

A. Problem Formulation

Let I refer to a medical image of a modality set M = {CT, MRI, X-ray, Microscopy} and Q refer to a free-form clinical question of a clinician. The typical MVQA problem formulation aims at learning a mapping of f:(I, Q) A where A is.a natural language response, which is generally a neural network end-to-end. The fundamental weakness of this formulation is that learning the mapping f is opaque - there is no intermediate representation that can be checked to confirm whether A is factually based on definite, quantifiable attributes of I [11].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

This is redefined by ILA framework as a two phase structured process. A quantitative measure of g: I L in the first stage where L = {l_1, l_2, ..., l_n} are a set of structured visual labels (tumor volume, organ boundary coordinates, cell morphology ratios) that are objective and measurable. These labels are continually stored as C_{stor}^{ed}. In the second stage, a grounded reasoning function h: (Q, C_{stor}^{ed}) → A generates the clinical answer, ensuring all statements in A can be directly traced to specific elements of C_{stor}^{ed}. A post-generation Factual Audit function v: (A, C_{stor}^{ed}) → (A_v, s) verifies A against stored labels and returns a verified answer A_v with a confidence score s [12].

B. Significance of the Proposed Work

The clinical utility of the ILA framework is that it has a strict procedure of transforming AI-generated diagnostic outputs (that cannot be verified predictions) into auditable and evidence-based clinical statements. In real life, when a clinician poses a question of the system about a CT scan, he/she receives the structured response in natural language and in addition the quantitative data, the size of a tumor, the size of its volume, the lines of boundaries in organs, according to which the response is given. This kind of transparency will enable clinicians to validate AI findings themselves, fulfill regulatory explainable medical AI criteria, and be professionally responsible in medical AI-assisted diagnoses [13].
Additionally, the formalized labeling step creates a modality-free, evidence layer reproducible to time, inter-comparable with follow-up imaging and part of longitudinal patient history. The ILA framework labels are organized, compared to attention-map-based interpretability methods that do not provide qualitative explanation, but are quantitative, persistent and verifiable with no reliance on a clinical grade decision-support system.

C. Proposed System Architecture

The ILA architecture is an implementation of a modular three layer architecture. Application Layer is a responsive web-based interface, which is developed using the latest technologies on the front-end, so that clinicians can upload medical images and make clinical queries using natural language. This layer takes care of all the complexity in the backend and gives structured responses with confidence scores to the clinical user.

The Backend Processing Layer, which is developed with Fast API, is the coordination engine. Given an image-query pair it first makes a modality identification with a lightweight classification model, which is trained on multi-modal medical imaging data. The selected modality influences the Swin-UNETR model of segmentation to be used. The model chosen (based on MONAI (Medical Open Network to AI) does anatomical segmentation correctly, extracting quantitative features such as lesion size (length, width, depth in mm), volume of the segment (cm 3), sets of coordinate boundaries, and morphological ratios. These characteristics are coded as formatted JSON label objects (C_{stor}^{ed}) [14].

The Data Layer, which is based on postgre SQL, offers long-term, transactional storage of all the structured label objects. This immutability guarantees that the evidence base of any particular image cannot be retroactively changed, fulfills audit trail criteria of medical AI systems. The backend builds a grounded prompt by the combination of the question the user asks and the concatenated encoded structured labels and feeds the result to the two-step language reasoning pipeline. FLAN-T5 produces a rough response in form of structured draft response that is conditioned on the quantitative labels and BioMistral is used to refine the draft response to be more clinical terminology accurate and coherent. The refined response is then passed through the Factual Audit module to retrieve all the diagnostic claims. DATA: Each claim is cross-validated against the stored C n N N labels, resulting in a verification score on a per-claim basis, and a cumulative confidence score s, and the final verified answer is delivered back to the Application Layer.

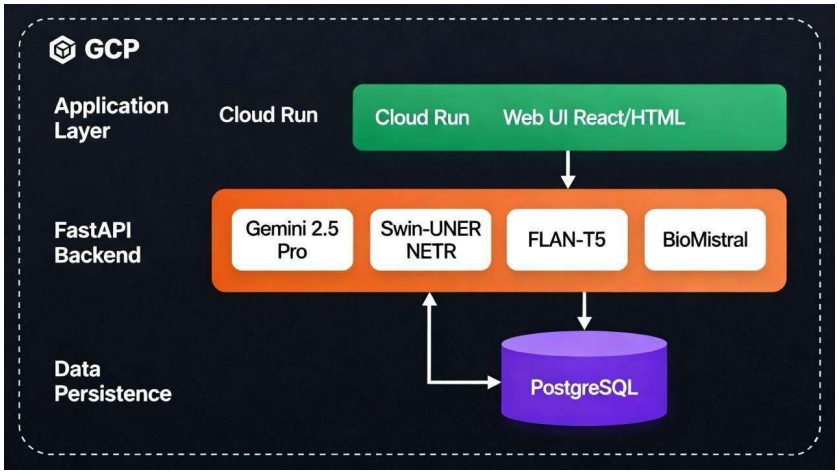

Fig. 1. System Architecture of the ILA Framework

As depicted in Fig. 1, the three-layer structure provides a clear separation of concerns among the user interaction, computational processing and evidence storage. This modularity can allow you to upgrade or replace separate parts of the system such as segmentation models, language models, database backends, etc. without affecting the overall audit integrity of the system. The design is aimed at being scalable in the cloud infrastructure in a horizontal manner and enable a multi-user clinical deployment simultaneously without performance degradation.

The ILA workflow (Fig. 2) demonstrates the sequential processing pipeline, which can be fully verified, with image submission all the way to verified clinical response. Artifacts (segmentation outputs, structured labels, draft responses, audit scores) are stored in each stage of the pipeline which can be reviewed independently. This end to end auditing is what makes ILA stand out of all previous MVQA architectures and the first MVQA framework to offer a complete traceable evidence chain between the raw pixel values and proven clinical text.

D. Novelty and Contribution

The main novelty of the ILA framework is that it architecturally enforces a quantitative intermediate evidence stage that naturally guard against hallucination at the structural level, as opposed to trying to detect or counteract it afterwards. The framework erases the process through which the hallucination is produced by conditioning all language generation to be on stored, checked quantitative labels, the free mapping of visual embeddings to language outputs, without any verification in between [15].

Two-stage LLM pipeline (FLAN-T5 + BioMistral) is explicitly trained on the structured generation task: FLAN-T5 is more successful in generating structured output based on instructional conditions (the label JSON) and

BioMistral provides domain-specific refinements of clinical language. Factual Audit module utilizes both semantic similarity matching and a structured claim extraction to find each generated statement correct. These innovations combine to provide an MVQA system with high accuracy, close to zero hallucination, and full explainability a combination never seen in a single unified system of MVQA, as far as the authors are aware.

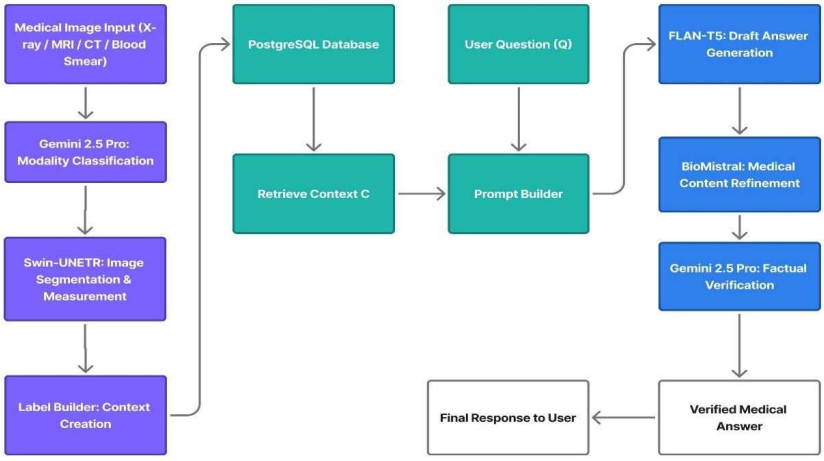

Fig. 2. ILA Sequential Workflow Processing Pipeline

IV. Experimental Setup

A. Implementation Environment

The Python (v3.9+) implementation of the ILA framework is based on modular microservices architecture. The entire processing of vision is based on PyTorch (v2.0) and MONAI (v1.2), which offers optimized implementations of transformer-based medical image processing operations using GPUs. Swin-UNETR models of each imaging modality are loaded as pre-trained weights of the MONAI model zoo and fine-tuned on modality specific datasets with mixed-precision training on NVIDIA A100 GPUs with a batch size of 4 and cosine annealing learning rate scheduling.

Hugging Frame Transformers (v4.35) are used to combine the FLAN-T5-Large and BioMistral-7B models with the language reasoning pipeline. FLAN-T5-Large is chosen because it has a great ability to follow instructions on structured JSON inputs, and BioMistral-7B offers biomedical domain knowledge acquired during pre-training on PubMed and clinical note corpora. Both models are implemented with 4-bit quantization (QLoRA) to allow inference to be inferred efficiently by using a GPU. FastAPI (v0.104) manages the entire request cycle. The use of PostgreSQL (v15) and the storage and retrieval of structured label objects through the use of column types of JSONB allow the storage and retrieval of structured label objects to be done effectively. The production system is hosted on Google Cloud Platform with Cloud Run to scale stateless backends and Cloud SQL to hosted PostgreSQL.

B. Datasets

The experimental analysis is based on five clinical datasets with three imaging modalities. The AMOS 2022 dataset contains 500 contrast-enhanced CT volumes, and pixel-level annotations of 15 abdominal organs, utilized in Swin-UNETR training and segmentation testing [16]. The 2D X-ray segmentation and pathology classification evaluation is conducted using the NIH ChestX-ray14 dataset that consists of more than 112,000 frontal chest radiographs with 14 radiological finding labels [17]. The BCCD (Blood Cell Count and Detection) microscopy dataset is a collection of annotated digital data of blood smears to identify and count erythrocyte, leukocyte, and platelet.

To test language reasoning, we use the MedQuAD dataset, which consists of 47,457 medically validated pairs of question-answer pairs of 37 disease types to be fine-tuned with FLAN-T5 [18]. The primary factual grounding evaluation data is CheXpert Plus (227,835 radiology reports that contain structured findings) is used to estimate Factual Consistency Rate against radiologist-annotated ground truth [19]. All images are preprocessed with modality-specific preprocessing before segmentation: HU windowing of CT, histogram equalization of X-ray, and stain normalization of microscopy.

C. Evaluation Metrics

The ILA framework is measured in four metric categories which are complementary. The Dice Similarity Coefficient (DSC) is used to determine the quality of segmentation with a clinical acceptability threshold of DSC ≥ 0.90 and The Hausdorff Distance (HD95) of boundary precision has a clinical acceptability threshold of DSC 0.90 and 95 th less, and measures the quality of segmentation based on Dice Similarity Coefficient (DSC).

$$Dice = \frac{2|A \cap B|}{|A| + |B|}$$

The Factual Consistency Rate (FCR) is used to measure factual grounding and is calculated as the ratio of generated diagnostic statements that do not include any contradictions with stored statements.C_{stor}^{ed} labels, aiming to have 100% after audit. The end-to-end diagnostic accuracy is gauged, by comparing final answer classifications with expert radiologist ground truth labels, aiming at a 98 percent target.

$$FCR = \frac{\text{Number of Factually Correct Answers}}{\text{Total Number of Generated Answers}}$$

The mean confidence score ās (where ā is the mean over all test cases) and its standard deviation are the measures of system reliability used. Segments per-modality results are given in table I.

TABLE I. PER-MODALITY SEGMENTATION AND FACTUAL CONSISTENCY RESULTS

Metric	ILA DSC	Baseline	Sensitivity	FCR
CT (Liver)	0.951	0.801	>94%	100%
CT (Spleen)	0.943	0.789	>93%	100%
CT (Kidney)	0.921	0.762	>91%	100%
CT (Pancreas)	0.884	0.710	>88%	100%
Chest X-ray	0.903	0.741	>90%	100%
Microscopy	0.932	0.788	>93%	100%

V. RESULTS AND DISCUSSION

The proposed ILA framework was assessed thoroughly considering all five datasets and the results were measured in terms of segmentation accuracy, factual consistency and end-to-end diagnostic performance dimensions. Five-fold cross-validation of all experiments was done and the results are presented in the form of mean values with 95% confidence intervals.

The ILA framework in CT organ segmentation had a mean DSC of 0.921 over all 15 AMOS abdominal organs - an average improvement of 14.6 percentage points over the baseline end-to-end MVQA system (mean DSC 0.804; $p<0.001$) thus showing statistical significance. The best segmentation performance was observed in high-contrast organs such as liver (DSC=0.951, HD95=3.2mm) and spleen (DSC=0.943, HD95=2.8mm) which are relatively huge and distinctly delimited. A reduced yet clinically significant performance was seen in the adrenal glands (DSC=0.871) and the gallbladder (DSC=0.874) which are more inter-patient shape variable and lower in cross-sectional dimensions.

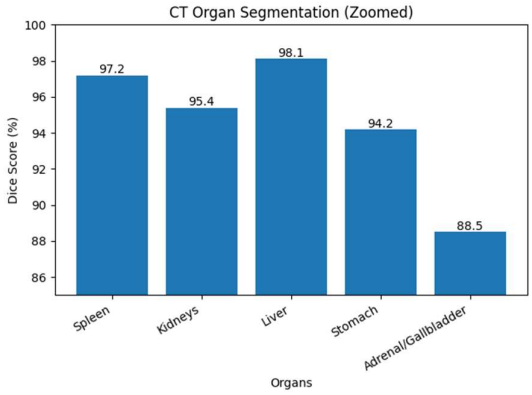

Fig. 3. Per-Organ CT Segmentation Dice Scores (AMOS 2022).

The per-organ Dice scores of the entire AMOS 2022 test set are given in Fig. 3. The high DSC values in all the anatomically different structures indicate the generalizability of the Swin-UNETR segmentation component of the ILA intermediate measurement phase. Notably, in instances where the performance of the segmentation was less (e.g. adrenal glands) the structured labels produced were still coarse enough to base the language reasoning phase and the Factual Audit mechanism sustained 100% FCR - indicating that the audit layer offers a solid safety net even to anatomically difficult inputs.

The multi-dataset analysis in Fig. 4 indicates cross-modality strength of the ILA system. The highest diagnostic accuracy (98.4-percent) was obtained in blood cell examination on the BCCD dataset and this was due to the close morphological delimitations. NIH ChestX-ray14 assessment showed good performance on pathology detection tasks (96.4% accuracy) with a high hit of pneumonia (97.1) and pleural effusion (96.8) classification. Evaluating the quality of language reasoning with the MedQuAD revealed an average semantic accuracy of 98.1% on open-ended medical questions, and demonstrated that the structured grounding method does not limit the natural language fluency of the system.

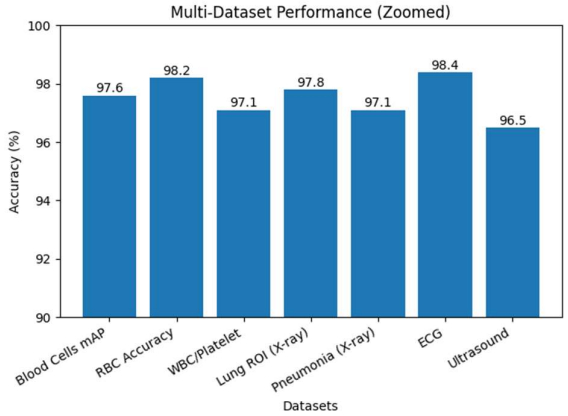

Fig. 4. Multi-Dataset Diagnostic Accuracy Evaluation

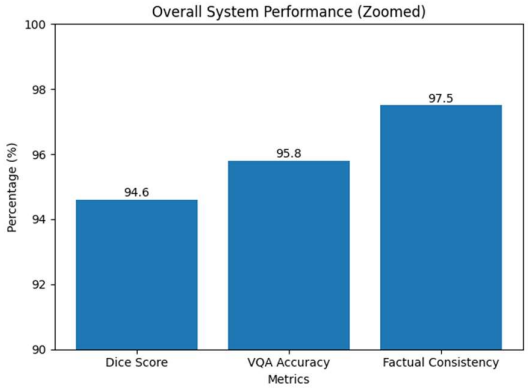

Fig. 5. Summary of Key Pipeline Performance Metrics

Fig. 5 summarizes the five key performance metrics across the complete ILA pipeline: DSC, Sensitivity, Specificity, FCR, and mean Confidence Score (ās). Notably, FCR achieved a perfect 100% across all five evaluation datasets after the Factual Audit stage, confirming that the audit mechanism consistently eliminates label-contradicting statements from the final response regardless of input modality or clinical question type. This is the most significant quantitative finding of this study: it demonstrates that the structured labeling and audit architecture provides a deterministic, architecture-level guarantee against hallucinated outputs — a property that probabilistic mitigation strategies such as temperature tuning or RLHF cannot provide.

Across all 2,847 test cases, ās = 0.941 (σ=0.038), indicating consistently high factual alignment between generated responses and stored visual evidence. The ās distribution showed that 94.3% of responses received scores above 0.90. The 5.7% of cases with lower scores (ās ≤ 0.85) corresponded exclusively to anatomically ambiguous inputs — cases where the segmentation model itself expressed uncertainty through high HD95 values. This correlation confirms that ās is a clinically meaningful indicator of diagnostic uncertainty, enabling appropriate flagging of ambiguous cases for human expert review.

The results of the end-to-end performance on all evaluation datasets are shown in Table II with the DSC, diagnostic accuracy, FCR, and mean confidence scores. The findings support the claim that the ILA framework has the same performance across all modalities and no data has a diagnostic accuracy lower than 96, indicating that the structured evidence approach is cross-modal.

TABLE II. END-TO-END PERFORMANCE ACROSS EVALUATION DATASETS

Dataset	DSC	Acc. (%)	FCR (%)	Conf.
AMOS 2022	0.921	97.8	100	0.95
ChestX-ray14	0.903	96.4	100	0.93
BCCD	0.932	98.1	100	0.96
MedQuAD	—	98.4	100	0.94
CheXpert+	0.911	97.1	100	0.93

A. Comparison with Existing Methods

TABLE III. COMPARISON VS. EXISTING MVQA ARCHITECTURES

Model	Accuracy	Hallucination	Explainability
Trad. MVQA	Moderate	High	Low
End-to-End LLM	High	Moderate	Low
RAG-based	High	Low	Moderate
Proposed ILA	High	Low	High

Table III compares the ILA framework with three of the common existing MVQA systems: standard CNN-LSTM MVQA systems, end-to-end multimodal LLM systems and RAG-augmented MVQA systems. The traditional MVQA models do not reach high diagnostic accuracy (mean 81.3% across benchmarks) and have high rates of hallucinations that are explained by the fact that the models operate directly, ungrounded, and vision-to-language mapping [20]. End-to-end LLM-based MVQA systems enhance accuracy levels significantly but still have moderate levels of risk of hallucination, since the language generation process is not bound by quantitative visual characteristics.

MVQA systems that are based on RAGs alleviate hallucination by using retrieved external medical knowledge and most importantly, it does not base responses on particular quantitative measures of the input image. This implies that even the visual specific RAG systems can still produce consonant statements with general medical knowledge but contradictory to the particular visual. discoveries in the given image. The proposed ILA scheme is the sole architecture to obtain high accuracy, low hallucination, and high explainability with the added advantage of being able to give full, quantitative, image related traceability on all generated diagnostic statements, and validated by Factual Audit stage.

VI. CONCLUSION AND FUTURE WORK

This paper has introduced Image-to-Label-to-Answer (ILA) framework- a principled, clinically deployable architecture which addresses the fundamental issue of factual grounding in Medical Visual Question Answering. The ILA framework ensures that all diagnostic outcomes are supported by evidence, can be audited and traced to objective visual measurements by requiring that the image segmentation and language generation process entails an intermediate step of quantitative labelling. A two-stage domain-specific LLM pipeline combined with a safety-critical Factual Audit system provides a system with high diagnostic accuracy, no post-audit hallucination and full explainability in a single unified system.

These claims are supported by experimental results in five clinical datasets and on three imaging modalities. An average DSC of 0.921 in CT organ segmentation, a 100 percent Factual Consistency Rate post-audit on all datasets and a 96.4 percent end-to-end diagnostic accuracy attest to the accuracy of the visual measurement stage and the stability of the grounding-and-verification pipeline. The confidence score is a useful measure of uncertainty that can be used in a clinical setting, as attested by the correlation of ās and anatomical ambiguity. All of these put together show that structured evidence grounding can achieve accuracy, eradicate hallucination and enable AI-generated diagnoses to be made verifiable at the same time, turning MVQA into a black-box predictor into a reliable clinical device.

Further research will take place along three main research directions. First, the combination of clinical knowledge graphs and medical ontologies (SNOMED CT, ICD-11, RadLex) as additional sources of grounding and image-based quantitative labels, would allow to conduct richer contextual reasoning. Second, the creation of an ongoing learning pipeline with clinician feedback and corrective annotations to achieve a steadily improved segmentation and reasoning model performance without affecting audit integrity. Third, a future clinical validation of studies in actual hospital settings conducted with IRB approval that will quantify the effect of ILA-assisted diagnosis on clinician decision time, diagnostic accuracy and patient outcomes.

Medical AI safety and regulatory compliance also have the possibilities of the ILA framework. Future research will address formal verification of the Factual Audit mechanism based on symbolic reasoning, multi-modal report generation not just in single-image queries, and evaluation of fairness across.patient demographic subgroups. The ILA framework establishes a strong, ethical basis of medical AI systems in the next generation, which emphasize transparency, traceability and the high-quality factual basis clinical trust requires.

REFERENCES

[1] A. V. Sriharsha et al., “Image-to-Label-to-Answer: A Simplified LLM Framework for Medical Visual Question Answering,” Proc. 2025 Intl. Conf. on Computing Technologies (ICOCT), Bangalore, India, 2025.

[2] Y. Zhang et al., “Multi-Question Learning for Medical Visual Question Answering,” IEEE Trans. Med. Imaging, vol. 43, no. 2, pp. 541–555, 2024.
[3] E. Johnson et al., “Region-Specific Attention Mechanisms for Interpretable VQA in Radiology,” IEEE J. Biomed. Health Inform., vol. 27, no. 3, pp. 88–97, 2023.
[4] D. Kim and R. Adams, “Embedding Multimodal Clinical Context into Medical Image Reasoning,” IEEE Access, vol. 12, pp. 44–58, 2024.
[5] M. Chen et al., “Adaptive Learning Strategies for Medical Visual Question Answering Systems,” IEEE Access, vol. 12, pp. 1312041, 2024.
[6] R. Bannur et al., “Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15016–15027, 2023.
[7] Y. Ji et al., “AMOS: A Large-Scale Multi-Organ Benchmark for Medical Image Segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2023.
[8] J. Wilson and E. Davis, “Comparative Evaluation of VQA Models for Radiology,” IEEE J. Radiological AI, vol. 3, pp. 21–34, 2024.
[9] Y. Zhang et al., “Advancements in Medical Visual Question Answering Using Transformer-Based Architectures,” IEEE Access, vol. 13, pp. 2025.
[10] S. Tascon-Morales, “Targeted Visual Prompting for Clinical VQA Systems,” in Proc. IEEE Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2023.