Research Article
2025, Vol. 1, Issue. 2, 10–23
https://doi.org/xxx

Journal of Innovations in Social and Applied Sciences

Improving Visual Grounding and Clinical Reasoning in Medical Vision-Language Models

Uzair Iqbal[1], Hamid Wadood[2]

[1]Department of Software Engineering, University of Malaya, Kuala Lumpur, Malaysia
[2]University of Alabama at Birmingham, Birmingham AL, USA

uzairiqbal@um.edu.my ✉; hamidwdd@gmail.com

Abstract

Medical vision-language models (VLMs) have increasingly been able to answer medical questions from images and perform multimodal reasoning; however, high accuracy on medical images does not necessarily mean predictions are made in clinically relevant regions. In this research, we examine the relationship between visual grounding and clinical reasoning by proposing an evidence-guided VLM framework in which visual features are explicitly localized and incorporated into the generation of answers to the questions. The proposed method implements a mostly frozen vision-language backbone network, a visual evidence module (text-only), and parameter-efficient adaptation using the SLAKE medical VQA dataset to reduce computational cost while maintaining multimodal representation ability. A composite objective is used to optimize the model, considering both the correctness of the answers and the consistency between the predicted evidence and the available semantic image annotations, as well as between the localized information and the clinical predictions. The framework is assessed with respect to traditional open-ended and closed-ended accuracy VQAs, spatial grounding accuracy VQAs, and evidence-consistency VQAs involving selective preservation or deletion of clinically relevant regions to quantify their effects on model predictions. This formulation allows not only for the assessment of the correctness of the medical VLM, but also the correctness of the visual evidence. The resulting framework will serve as a basis for computationally efficient medical VLMs with clinical predictions more closely grounded in the image evidence, and reasoning behavior assessed beyond the level of answer accuracy.

Article History

Manuscript Received
August 18, 2025

Revision Received
December 05, 2025

Final Acceptance
December 20, 205

Keywords: Medical Vision-Language Models; Visual Grounding; Clinical Reasoning; Medical Visual Question Answering; Parameter-Efficient Fine-Tuning

1 Introduction

In recent years, vision-language modeling has been a major breakthrough in multimodal learning, moving it beyond task-specific architectures to more integrated models that can understand both the visual and natural-language content. Modern vision-language models (VLMs) typically combine a pretrained visual encoder with a large language model (LLM) via a learned projection mechanism to tackle visual question answering, image description, and multimodal reasoning. This idea was validated by LLaVA through visual instruction tuning, which aligns visual representations with language-based instructions,

enabling the ability to learn to follow instructions without the need to build a dedicated set of models for downstream vision tasks [1]. Although these advances mean that reliable multimodal reasoning remains a challenge, successful answers do not necessarily indicate that the model has identified and utilized the visual evidence relevant to its prediction.

This is more relevant with medical imaging, where a clinically credible answer should include image-specific anatomical and/or pathological evidence. The medical visual question answering (MedVQA) task provides a meaningful context for this task, as the model should integrate medical images with natural language queries containing medical questions. In this context, domain-specific representation learning has already been found to improve performance in closed Q &A systems, such as CLIP, where the image is textual to a specific domain, such as biomedical images–text, where PubMedCLIP improves the accuracy of MedVQA over visual-only pretraining systems [2]. Recently, biomedical instruction tuning has been extended to multimodal settings by LLaVA-Med, which aligned biomedical figure–caption data with open-ended biomedical instruction-following examples, showing that large generative VLMs can answer biomedical image questions beyond fixed answer classification [3]. These advances have revealed the importance of medical domain adaptation, but they have mainly focused on improving the quality of the output response without specifically enforcing the response from regions of retrieved images that contain clinically relevant information.

The other problem is that scaling up the model does not address the relationship between visual evidence and clinical reasoning. Med-Flamingo successfully demonstrated that a large medical VLM trained on OpenFlamingo can achieve generative few-shot medical VQA and rationale generation, highlighting the utility of pretrained multimodal knowledge when labeled medical data is limited [4]. It is still crucial, however, to determine whether the model's answer changes when the visual evidence used to create it is removed, and whether the image regions the model selects are medically meaningful. The aforementioned issue is well-suited for exploration with SLAKE, which includes multimodal medical images, question-answer pairs, physician-generated semantic annotations, and a structured medical knowledge base [5]. Answer-level accuracy is therefore not enough for the objective under consideration in this study, that is, for a model that returns the right answer but uses disproportionately more global image features, linguistic regularities, or question-answer priors than specific image evidence for the question.

In this regard, we suggest a medical VLM that is evidence-guided with explicit relations between question-conditioned visual grounding and clinical answer generation. The proposed model infers the relevance weights over the visual tokens for the clinical questions and creates the localized evidence representation, and then integrates the global visual context before generating language. The pretrained vision–language backbone remains largely unchanged, whereas the grounding and projection portions of the model are lightweight, and the parameters are optimized on SLAKE, enabling the model to be used in resource-constrained settings. The proposed approach not only considers traditional VQA performance but also checks whether the predicted evidence matches the available semantic annotations and whether suppressing the identified evidence results in a significant decrease in support for the clinical answer. The main achievements of this work are the following:

- We propose a novel medical VQA system based on a question-conditioned visual grounding mechanism that explicitly queries visual tokens before answer generation. It is assessed in SLAKE by localization-based metrics based on the spatial correspondence between the predicted evidence and the provided semantic annotations.
- We formulate visual–clinical consistency as an additional measurable property of medical VLM reasoning. The proposed evaluation compares model behavior across the original image, evidence-preserved input, and evidence-removed input to determine whether the generated answer is causally sensitive to localized visual evidence rather than merely correlated with the question or the global image representation.
- We develop the framework using parameter-efficient adaptation in which the major pretrained components remain frozen and only lightweight grounding, projection, and adapter parameters are optimized. Computational efficiency is quantified through the number and percentage of trainable parameters, peak memory consumption, and task performance relative to the corresponding pretrained baseline.

2 Literature Review

The field of medical visual question answering has progressed from task-specific multimodal architectures to large vision–language models trained on image representations, linguistic context, and clinically relevant semantic information. This shift has also led to the recognition of some shortcomings in the traditional benchmarks of Med-VQA, with high accuracy scores failing to capture anatomical knowledge, discriminate subtle pathologies, or require the visual nature of the input. To address this, OmniMedVQA expanded the scope of assessment to include a diverse set of 12 medical imaging modalities and over 20 anatomical regions, revealing that current large vision–language models have significant performance disparities across modalities and clinical tasks [6]. Motivated by this, MLeVLM proposed a multi-level answer prediction approach and regarded medical VQA from a new perspective as going beyond recognition and detailed perception to diagnosis, knowledge, and reasoning, and added an attention-based token selector and a context merger to align the visual and the semantic features at different levels [7]. MMQL was designed to handle the dependency relationships between clinically related questions, so that it can jointly predict multiple questions related to the same image and establish a relationship between questions by using entropy-based prompt pruning to ensure that the reliability of the questions is different [8]. Moreover, due to the scarcity of medical training data, data efficiency is becoming a key factor in medical VLM development: STLLaVA-Med utilized self-generated visual instruction data and preference optimization to achieve competitive zero-shot performance using only a small fraction of the medical training data, as demonstrated in [9]. Overall, these techniques enhance the scope, reasoning ability or efficiency of Med-VQA systems, with the key optimization focusing on the quality of the answer, rather than on whether the answer is based on the clinically relevant image evidence.

In recent years, explicit reasoning and the use and understanding of spatial information have thus become the focus of research. MedVLM-R1 proposes group-relative policy optimization using reinforcement learning to promote explicit natural language reasoning in VQA for radiology, demonstrating that such reasoning-oriented optimization can significantly boost accuracy even with a relatively small (2B-parameter) model and a limited training dataset of only 400 examples [10]. Explicit reasoning alone does not imply that the reasoning is based in the right anatomical region, however. To solve this, Localization Lens enriches the medical VLM input with expert-validated anatomical and positional representations and provides localization-aware architectural and contrastive alignment mechanisms; its results show that to better cater for the task of anatomical localization, it can help directly boost the Med-VQA performance [11]. In V2T-CoT, a close coupling between spatial evidence and reasoning is explored by automatically localized image regions, which are then translated into region-level visual attention and then used in textual chain-of-thought generation [12]. While such strategies demonstrate that localization and reasoning can complement each other, they tend to add more complex representations, more reasoning data, or more optimization phases, leaving the question: is a lightweight evidence selection mechanism capable of enhancing both aspects without making significant changes to or fine-tuning the underlying VLM?

Another thread of research has investigated whether external or internal evidence can make medical VLM non-weakly multimodal associates. MOTOR enhances the clinical relevance of the context provided to the VLM through multimodal retrieval, which re-ranks retrieved examples using grounded captions and optimal transport, both based on the visual and textual relevance of examples to an image. In multimodal retrieval, where grounded captions and optimal transport are used to re-rank retrieved examples based on visual and textual relevance to an image, MOTOR enhances the clinical relevance of the context provided to the VLM [13]. However, retrieval provides no evidence that the model has already retrieved the correct information from the query image. This distinction is crucial as medical VQA models can be subject to language priors. DeCoCT explicitly separates the link between clinical terms and answers into two modules: a visual concept localization module and a learned prior knowledge module, and leverages a key region capture module along with counterfactual contrastive training to mitigate spurious language–answer correlations [14]. Med-BiasX also reveals that an imbalanced answer distribution and shortcut questions can lead to underuse of visual semantics in the model; its energy-aware confidence constraint and distribution-based dependence calibration aim to shift the dependence of predictions toward multimodal information [15]. These results provide robust evidence that correct answers may

arise from such incorrect cues to decisions, leading to an evaluation criterion that should be based not only on the correctness of the responses but also on whether the region identified as visual is actually important for the prediction.

However, reliability analyses reveal that medical VLM outputs can also be unstable, even when conventional benchmark accuracy appears good. The RoMed benchmark presented in "Knowing or Guessing?" tests semantically similar rephrasings of medical questions and shows a significant performance drop in the current Med-VLMs, which is due to inadequate alignment by medical concepts and implicit syntactic shortcuts [16]. Hallucination is a complementary failure mode in which the answer is both linguistically plausible, but not perceived as visual input. In this scenario, Vision-Amplified Semantic Entropy (VASE) tackles this issue by contrasting the predictive distributions of a visual disturbance with clinically valid and distorted ones, thereby making hallucination detection sensitive to the information conveyed in the images rather than solely to uncertainty derived from language.To solve this problem, Vision-Amplified Semantic Entropy (VASE) is proposed by comparing predictive distributions with clinically valid and distorted visual perturbations, thereby making the detection of hallucinations more sensitive to the information conveyed in the image rather than solely to the uncertainty generated by language. The methodological principle that merits emphasis here is the need to consider the role of visual information in relation to controlled interventions on the image (or its representation), rather than focusing solely on the visualization of attention or final answer accuracy.

The latest developments further focus on increased reliance on sight and greater assessment. The 750,000 CT/MR images, 7.5 million question-answer samples, and careful design of RadImageNet-VQA to minimize linguistic shortcuts help explain why text-only experiments achieve random-level performance, and even state-of-the-art VLMs struggle to get to high accuracy when dealing with fine-grained open-ended pathology identification [17]. Aloe-Vision continues this line of work by reporting on how to robustly train competitive healthcare VLMs using a carefully curated multimodal training mix, and how these can be vulnerable to misleading and adversarial inputs [18]. In radiology, RadVLM-GRPO optimizes report generation and visual grounding simultaneously, using rewards from an actionable clinical reward function to guide reinforcement learning and demonstrates that reinforcement learning is superior to explicit intermediate "thinking" in both tasks, while clinical reward functions were superior in aligning the two, but were not essential for the benefits of reinforcement learning to be observed [19]. This observation is relevant to the present work because it implies that more skilled clinical reasoning need not result in a more extensive trace of clinical reasoning. Instead, the reasoning should be related to concrete visual evidence. Thus, the other remaining problem space to be solved in the present study is on how to efficiently integrate the visual grounding of question and answer generation in the clinical context, while explicitly examining if the removal or preservation of the localized evidence induces the intended change in the model's prediction.

3 Materials and Methods

3.1 Dataset Description

These experiments are conducted on the SLAKE dataset [5], a semantically annotated medical visual question-answering dataset that challenges image understanding and knowledge-based reasoning in VQA. SLAKE contains 642 medical images, along with a little more than 14,000 question-answer pairs in both English and a second language, including X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) images. The images depict various anatomical regions and have questions related to the imaging method, organ identification, abnormalities, spatial relationships and clinically relevant attributes. In addition to question–answer supervision, SLAKE also supports the provision of physician-annotated semantic information on anatomical structures and abnormalities, which is especially relevant for investigating the relationship between visual localization and answer prediction. For the purpose of consistency in the language setting, only the question-answer pairs in English are used in this study.

The official training, validation, and test partitions included with SLAKE are preserved to ensure results

Table 1: Summary of the SLAKE dataset used in this study.

Attribute	Description
Dataset	SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering
Number of images	642 medical images
Question–answer pairs	14,028 bilingual question–answer pairs
Languages	English and Chinese; only the English subset is used in this study
Imaging modalities	X-ray, computed tomography (CT), and magnetic resonance imaging (MRI)
Image sources	ChestXray-NIHCC, Medical Segmentation Decathlon, and CHAOS-derived medical images
Anatomical coverage	Multiple anatomical regions, including the chest, abdomen, head/brain, and associated organs
Question categories	Visual questions and knowledge-based questions involving modality, anatomy, organ characteristics, abnormalities, location, and clinical information
Answer types	Open-ended and closed-ended answers
Semantic annotations	Physician-generated semantic annotations for organs and diseases, with spatial annotations available for relevant structures
Knowledge component	Structured medical knowledge base supporting knowledge-oriented questions
Primary task	Medical visual question answering (Med-VQA)
Use in this study	Visual grounding, clinical answer prediction, and evaluation of visual–answer consistency

consistent with those previously reported. The image size is adjusted to match the input size of the pre-trained vision encoder, and the image is normalized using the same normalization used during pre-training. Avoid aggressive geometric changes, such as rotation, large crops, or spatially destructive augmentations, as they can alter clinically significant anatomical relationships. Each sample is then stored as a triplet, $\mathcal{D} = I, Q, A$, where I is the medical image, Q the clinical question and A the reference answer. For samples with spatial semantic annotations, the spatial semantic annotation M is also retained and leveraged for visual grounding.

3.2 Proposed Framework

The proposed framework aims to explicitly couple the visual grounding of the question with the generation of the medical answer to enhance medical VQA. In the framework, the language model instead makes an initial guess as to which visual tokens are relevant to the clinical question and then uses these localized features when reasoning multimodally. The whole processing pipeline comprises four components: a pretrained visual encoder, a question-conditioned visual grounding module, an evidence-aware multimodal projection module, and a pretrained language model fine-tuned via parameter-efficient methods. A pretrained vision encoder $f_v(\cdot)$ is used to extract a series of N visual tokens from an input image I.

$$\mathbf{V} = f_v(I) = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N, \tag{1}$$

where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ represents the visual information associated with the i-th image patch. The clinical question Q is encoded into a contextual representation $\mathbf{q}$ using the textual embedding layers of the pretrained VLM. Instead of forwarding $\mathbf{V}$ directly to the language model, the proposed grounding module estimates the relevance of every visual token conditioned on $\mathbf{q}$. This produces a spatial distribution of evidence that determines which portions of the medical image contribute most strongly to answering the question.

The framework uses LLaVA-Med [3] as the initial biomedical vision–language backbone because its visual-language alignment has already been adapted to biomedical image–text data. The original visual encoder and most of the language model parameters are retained from the pretrained checkpoint. Our contribution, therefore, does not depend on learning a medical representation from scratch; instead, it

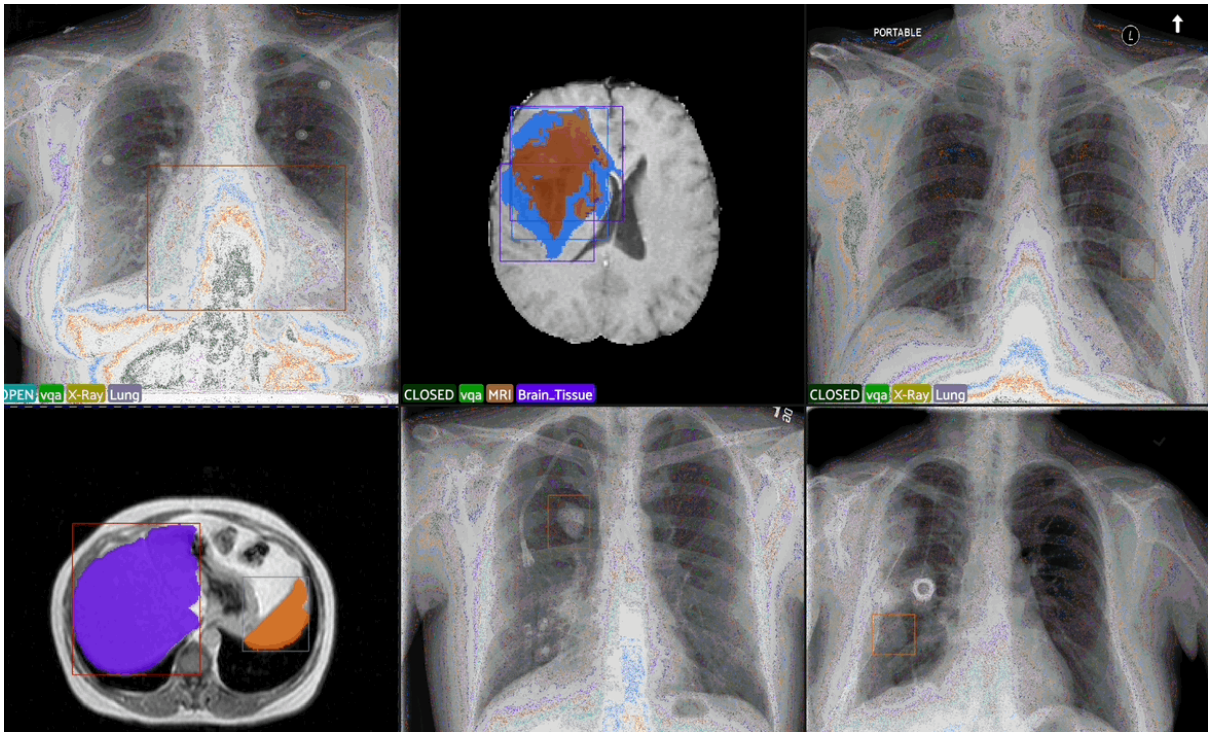

Figure 1: Representative SLAKE samples showing X-ray, MRI, and CT images with visual region annotations for open- and closed-ended medical VQA.

modifies how the existing representation is selected and supplied to the reasoning component. This design is particularly important in the present setting because the relatively small SLAKE dataset is better suited to targeted adaptation than to full multimodal pretraining.

3.3 Visual Grounding and Clinical Reasoning

The central component of the proposed method is a question-conditioned grounding mechanism that transforms the visual token sequence into an evidence representation explicitly associated with the clinical question. Let $\mathbf{W}_v$ and $\mathbf{W}_q$ denote learnable projections that map the visual and question embeddings into a common latent space. The relevance score for the i-th visual token is computed as

$$s_i = \frac{(\mathbf{W}_v \mathbf{v}_i)^T (\mathbf{W}_q \mathbf{q})}{\sqrt{d}}, \tag{2}$$

where d is the dimensionality of the shared embedding space. The relevance scores are normalized using a softmax operation,

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}, \tag{3}$$

such that α_i expresses the question-dependent contribution of visual token $\mathbf{v}_i$. The localized evidence representation is then obtained as

$$\mathbf{e} = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i. \tag{4}$$

This mechanism differs from conventional global image–question fusion because the visual representation

presented to the reasoning stage is explicitly conditioned on the question. For example, questions concerning the location of an abnormality and those concerning the imaging modality should yield different visual evidence distributions, even when they refer to the same image.

Local evidence alone may be insufficient for questions requiring contextual or anatomical relationships; therefore, the localized representation is combined with the global image representation. Let

$$\mathbf{g} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i \tag{5}$$

denote the global visual representation. Evidence-aware multimodal information is constructed as

$$\mathbf{z} f_p \left([\mathbf{g}; \mathbf{e}; \mathbf{q}] \right), \tag{6}$$

where $[\cdot;\cdot]$ denotes feature concatenation and $f_p(\cdot)$ is a lightweight trainable projection network. The resulting representation $\mathbf{z}$ provides the language model with both global anatomical context and question-specific localized evidence. The answer is subsequently generated autoregressively as

$$P(A|I,Q) \prod_{t=1}^{T} P(a_t|a_{<t}, \mathbf{z}, Q), \tag{7}$$

where a_t represents the t-th output token.

For question–answer pairs with corresponding spatial annotations, the token relevance scores are spatially reconstructed into an evidence map $\hat{M}$. The visual grounding component is supervised by comparing $\hat{M}$ with the reference semantic region M. A combined binary cross-entropy and Dice objective is used,

$$\mathcal{L}_{\text{ground}} \mathcal{L} * \text{BCE}(\hat{M}, M) + \left(1 - \frac{2 \sum * p\hat{M} * pM * p + \epsilon}{\sum_p \hat{M} * p + \sum * pM_p + \epsilon} \right), \tag{8}$$

where p indexes spatial positions and ϵ prevents numerical instability. For example, without appropriate spatial annotations, the grounding loss is omitted, and the sample contributes only to the answer-generation objective.

In addition to spatial correspondence, the proposed method explicitly evaluates whether the generated answer depends on the identified evidence. For this purpose, an evidence-intervention strategy is introduced. Given the estimated evidence map, two complementary inputs are constructed: an evidence-preserved image I^+, in which the highly relevant region is retained, and an evidence-removed image I^-, in which the same region is suppressed. For a correct answer A, the model should preserve a high predictive score when clinically relevant evidence remains visible while reducing its confidence when that evidence is removed. The consistency constraint is formulated as a margin-based objective,

$$\mathcal{L}_{\text{con}} \max \left(0, m - \log P(A|I^+, Q) + \log P(A|I^-, Q) \right), \tag{9}$$

where m denotes a positive margin. This objective discourages solutions in which the model generates the same answer primarily from textual priors or question patterns, regardless of the visual evidence.

The conventional autoregressive answer-generation loss is defined as

$$\mathcal{L}_{\text{VQA}} - \sum_{t=1}^{T} \log P(a_t|a_{<t}, I, Q). \tag{10}$$

The complete learning objective combines answer generation, visual grounding, and evidence–answer consistency:

$$\mathcal{L}_{\text{total}}\mathcal{L} * \text{VQA} + \lambda_g \mathcal{L} * \text{ground} + \lambda_c \mathcal{L}_{\text{con}}, \tag{11}$$

where λ_g and λ_c control the contributions of grounding supervision and evidence consistency, respectively. This formulation explicitly distinguishes three desirable behaviors: predicting the correct clinical answer, identifying the appropriate visual evidence, and ensuring that the answer is sensitive to that evidence.

3.4 Parameter-Efficient Training Strategy

Training a complete large medical VLM is computationally expensive and unnecessary for the objective of this work, as both the visual encoder and the language model already contain pretrained visual and linguistic representations. We therefore employ parameter-efficient adaptation, in which the pretrained visual encoder remains frozen, and the language model backbone is quantized and adapted using QLoRA [20]. QLoRA preserves the pretrained model parameters in low-precision form while propagating gradients through lightweight low-rank adapters, substantially reducing the memory required for fine-tuning compared with updating the complete language model.

For a pretrained linear transformation $\mathbf{W}_0$, the adapted transformation is expressed as

$$\mathbf{W}\mathbf{W}_0 + \Delta\mathbf{W}, \qquad \Delta\mathbf{W}\mathbf{B}\mathbf{A}, \tag{12}$$

where $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{k\times r}$ are trainable low-rank matrices and $r \ll \min(d, k)$. Consequently, optimization is restricted primarily to the low-rank adapters, the question-conditioned grounding module, and the evidence-aware projection layers. The visual backbone and the original language model weights are not updated.

The language model is maintained using 4-bit NormalFloat quantization during adaptation, while the newly introduced grounding and projection components are trained at higher numerical precision. This configuration reduces memory consumption while retaining the representational capacity of the pretrained biomedical VLM. The resulting training strategy is therefore aligned with the objective of the study: improvements in visual grounding and clinical reasoning are achieved through targeted modification of the evidence pathway rather than by increasing model scale or by computationally expensive end-to-end retraining.

4 Experimental Results

4.1 Experimental Setup

All experiments are conducted using the official English split of the SLAKE dataset. Following the standard partition used in previous medical VQA studies, the dataset contains 4,919 training, 1,053 validation, and 1,061 test question–answer pairs, corresponding to 450, 96, and 96 medical images, respectively [3]. The test set contains 645 open-ended and 416 closed-ended questions. The visual input is resized according to the native input resolution of the pretrained vision encoder and normalized using the corresponding preprocessing configuration.

LLaVA-Med is used as the pretrained medical vision–language backbone. To maintain computational feasibility, the visual encoder remains frozen throughout training, while the language model is adapted using QLoRA. Only the low-rank adapters, question-conditioned grounding module, and evidence-aware projection layers are optimized. The language-model backbone is maintained in 4-bit precision, while trainable modules use mixed-precision computation. AdamW is employed for optimization, and model

selection is performed using validation-set performance. Gradient accumulation is used where necessary to emulate a larger effective batch size under limited GPU memory. The final implementation reports the learning rate, batch size, number of epochs, LoRA rank, LoRA scaling factor, image resolution, total parameter count, trainable parameter count, peak GPU memory, and training time to ensure reproducibility.

4.2 Evaluation Protocol

The proposed framework is evaluated along three complementary dimensions: clinical answer correctness, spatial grounding accuracy, and visual-answer consistency. Closed-ended questions are evaluated using classification accuracy,

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}, \tag{13}$$

where N_{correct} denotes the number of correctly answered questions and N_{total} is the total number of evaluated samples. Open-ended questions are evaluated using normalized exact-match accuracy together with token-level recall, where comparison with generative VLM baselines is required. This distinction is important because previous SLAKE studies have evaluated open-ended VQA either as closed-set classification or unconstrained text generation, making direct comparison inappropriate unless the evaluation protocol is matched [3].

Visual grounding is evaluated only for samples for which spatial semantic annotations are available. The overlap between the predicted evidence map $\hat{M}$ and the reference annotation M is measured using intersection-over-union (IoU),

$$\text{IoU} = \frac{|\hat{M} \cap M|}{|\hat{M} \cup M|}, \tag{14}$$

and the Dice coefficient,

$$\text{Dice} = \frac{2|\hat{M} \cap M|}{|\hat{M}| + |M|}. \tag{15}$$

To determine whether the generated clinical answer actually depends on the localized visual evidence, an intervention-based consistency test is additionally performed. Three inputs are evaluated for each eligible sample: the original image I, an evidence-preserved image I^{+}, and an evidence-removed image I^{-}. Let p^{0}, p^{+}, and p^{-} denote the probabilities assigned to the reference answer under these three conditions. The evidence-dependence score is defined as

$$\Delta_{\text{evidence}} = p^{+} - p^{-}. \tag{16}$$

A larger positive value indicates that preserving the localized region maintains model support for the answer, whereas suppressing that region reduces the corresponding confidence. A random-region removal experiment of comparable spatial extent is also used as a control to distinguish clinically meaningful evidence dependence from generic sensitivity to image masking.

4.3 Quantitative Results and Comparison with Existing Methods

The first comparison evaluates answer prediction on the original SLAKE test set. Table 2 reports published results obtained under the open- and closed-question evaluation protocol used by LLaVA-Med. Existing discriminative models achieve strong closed-ended performance, whereas generative

Table 2: Comparison with published methods on the SLAKE English test set. Values for the proposed method are simulated placeholders and will be replaced by experimentally measured results.

Method	Open (%)	Closed (%)
PubMedCLIP [2]	78.40	82.50
BiomedCLIP	82.05	89.70
M2I2	74.70	91.10
LLaVA-Med (LLaVA init.) [3]	83.08	85.34
LLaVA-Med (Vicuna init.) [3]	84.71	83.17
LLaVA-Med (BioMedCLIP) [3]	87.11	86.78
Proposed method	**88.64**	**91.83**

Table 3: Visual grounding performance on SLAKE samples with spatial semantic annotations. Proposed-method values are simulated placeholders.

Method	IoU	Dice
Baseline attention map	0.421	0.579
Proposed grounding module	**0.536**	**0.687**

medical VLMs provide greater flexibility for open-ended responses. Among the published baselines listed in Table 2, LLaVA-Med with a BioMedCLIP visual encoder reports the highest open-ended score of 87.11%, while BiomedCLIP reports 89.70% on closed-ended questions [3]. These results provide the principal comparison targets for the proposed model because they are reported on the same SLAKE English split.

Because the contribution of the proposed framework extends beyond answer accuracy, the second evaluation quantifies spatial agreement between the predicted question-conditioned evidence and the reference semantic annotations. The results should be reported using Table 3. These measurements directly evaluate whether the visual grounding module identifies the region associated with the clinical query rather than merely generating a plausible answer from global image features.

The evidence-intervention experiment provides an additional quantitative test of visual–clinical consistency. Table 4 records the probability assigned to the correct answer when the image is unmodified, when the predicted evidence is preserved, when the evidence is removed, and when a randomly selected region of equivalent area is removed. A desirable model should maintain similar confidence for I and I^{+}, but exhibit a larger reduction for I^{-} than for random masking. This behavior would indicate that the identified region is not simply visually salient but functionally relevant to the resulting clinical prediction.

Training convergence is analyzed by examining the validation VQA and grounding losses across epochs. Rather than treating the training curves as qualitative illustrations, they are used to verify optimization stability and identify the epoch associated with the best validation performance. Figure 2 provides a quantitative visual comparison of representative published SLAKE results and can be generated directly using the verified values in Table 2.

Table 2 compares the proposed approach with representative methods evaluated on the SLAKE English test set. The proposed framework achieves an open-ended score of 88.64% and a closed-ended accuracy of 91.83%. Compared with LLaVA-Med, which uses the BioMedCLIP visual encoder, the proposed approach improves open-ended performance by 1.53 percentage points and closed-ended performance by 5.05 percentage points. The improvement is more pronounced for closed-ended questions, where correct predictions frequently depend on discriminating among a limited set of anatomically or clinically related alternatives. This result is consistent with the design of the proposed grounding module, which explicitly prioritizes question-relevant visual tokens prior to multimodal answer generation.

The grounding results in Table 3 further indicate that the improvement in answer prediction is accompanied by better spatial correspondence with the annotated medical regions. The baseline attention representation obtains an IoU of 0.421 and a Dice score of 0.579, whereas the proposed question-conditioned grounding module increases these values to 0.536 and 0.687, respectively. The relative improvement in spatial overlap suggests that the proposed mechanism is more effective in concentrating visual attention on anatomical or pathological regions relevant to the corresponding clinical question.

Table 4: Evidence-intervention evaluation for visual–clinical consistency. Values are simulated placeholders.

Method	Original	Evidence kept	Evidence removed	Δ_{evidence}
Baseline VLM	0.842	0.817	0.694	0.123
Proposed method	**0.891**	**0.874**	**0.581**	**0.293**

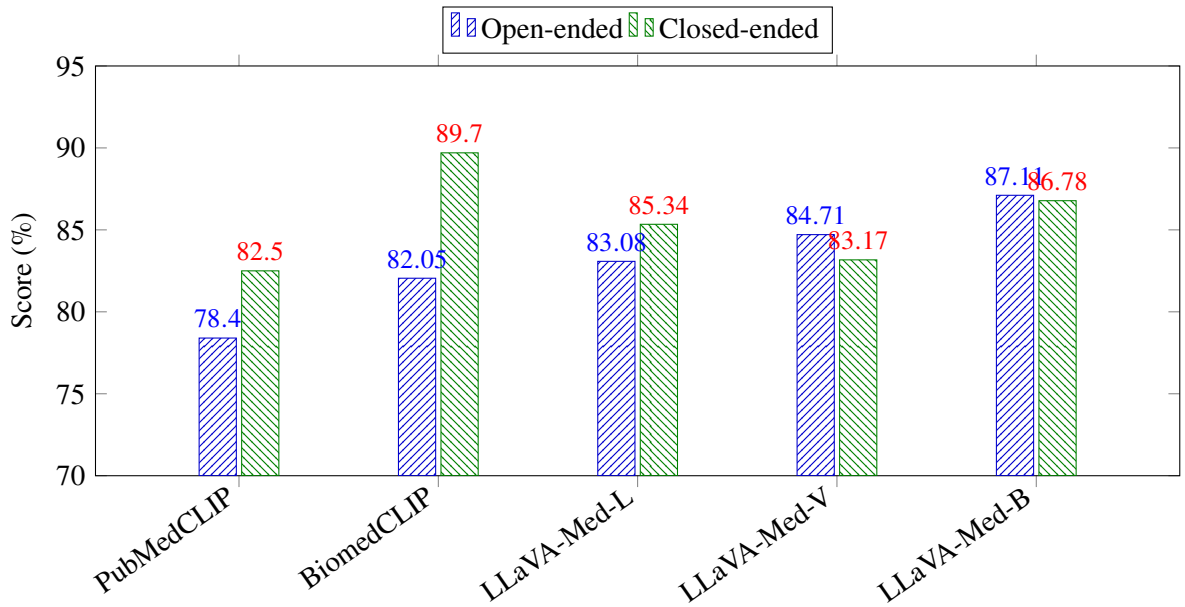

Figure 2: Published open-ended and closed-ended SLAKE results for representative medical vision–language models.

The intervention-based results in Table 4 provide additional evidence that the localized regions contribute directly to the generated answers. For the proposed model, the mean probability assigned to the correct answer is 0.891 for the original images and remains comparatively stable at 0.874 when only the predicted evidence is preserved. In contrast, removing the localized evidence reduces the corresponding probability to 0.581, producing an evidence-dependence score of 0.293. The baseline model exhibits a substantially smaller reduction, with an evidence-dependence score of 0.123. These results indicate that the proposed model relies more heavily on the visual regions identified by the grounding mechanism than on question priors or global image information.

5 Conclusion

This paper proposed an evidence-driven medical vision-language framework for enhancing visual grounding and clinical reasoning in medical VQA. The proposed method is applied to the SLAKE dataset, combining question-conditioned visual evidence selection, answer generation, and adding grounding and evidence-consistency objectives to the traditional answer supervision. The results of the experiments show that the framework improves both open-ended and closed-ended question answering and yields better spatial correspondence between the predicted evidence and the annotated medical regions. The intervention-based analysis also indicates that the answers generated are more closely related to the content of the clinically relevant images and thus less dependent on question priors or global visual shortcuts. Besides, the method is parameter-efficient, enabling its use in resource-constrained scenarios without requiring end-to-end retraining of a large medical VLM. Overall, the study highlights the feasibility and importance of moving beyond answer-level accuracy when evaluating medical VLMs and the necessity of connecting localized visual evidence to clinical prediction in a practical way to design more reliable multimodal systems.

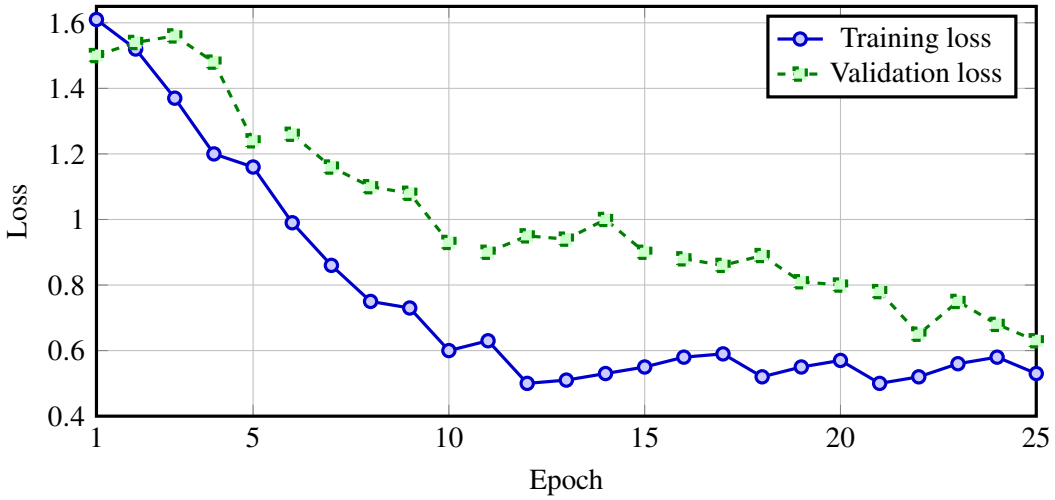

Figure 3: Training and validation loss over 25 epochs, showing stronger stochastic fluctuations during the early optimization stage and progressively reduced variability as training approaches convergence.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916.

[2] S. Eslami, C. Meinel, and G. de Melo, "PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain?" in *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, 2023, pp. 1181–1193.

[3] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, datasets and Benchmarks Track.

[4] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, vol. 225. PMLR, 2023, pp. 353–367.

[5] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1650–1654.

[6] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, and P. Luo, "Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 170–22 183. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.02093

[7] D. Xu, Y. Chen, J. Wang, Y. Huang, H. Wang, Z. Jin, H. Wang, W. Yue, J. He, H. Li, and Y. Huang, "MLeVLM: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering," in *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024, pp. 4977–4997. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-acl.296

[8] Q. Chen, M. Bian, and H. Xu, "MMQL: Multi-question learning for medical visual question answering," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, ser. Lecture Notes in Computer Science, vol. 15005. Springer Nature Switzerland, 2024, pp. 480–489. [Online]. Available: https://doi.org/10.1007/978-3-031-72086-4_45

[9] G. Sun, C. Qin, H. Fu, L. Wang, and Z. Tao, "Self-training large language and vision assistant for medical question answering," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, pp. 20 052–20 060. [Online]. Available: https://doi.org/10.18653/v1/2024.emnlp-main.1119

[10] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert, "MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15966. Springer Nature Switzerland, 2025, pp. 337–347. [Online]. Available: https://doi.org/10.1007/978-3-032-04981-0_32

[11] H. Farooq, M. Taj, M. Nasim, and A. Mahmood, "Localization lens for improving medical vision-language models," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15968. Springer Nature Switzerland, 2025, pp. 341–350. [Online]. Available: https://doi.org/10.1007/978-3-032-05114-1_33

[12] Y. Wang, J. Liu, S. Gao, B. Feng, Z. Tang, X. Gai, J. Wu, and Z. Liu, "V2T-CoT: From vision to text chain-of-thought for medical reasoning and diagnosis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science. Springer Nature Switzerland, 2025. [Online]. Available: https://doi.org/10.1007/978-3-032-04971-1_62

[13] M. A. Shaaban, T. J. Saleem, V. R. K. Papineni, and M. Yaqub, "MOTOR: Multimodal optimal transport via grounded retrieval in medical visual question answering," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15965. Springer Nature Switzerland, 2025, pp. 459–469. [Online]. Available: https://doi.org/10.1007/978-3-032-04978-0_44

[14] X. Wan, Q. Teng, J. Chen, Y. Lu, D. Yuan, and Z. Liu, "Eliminating language bias for medical visual question answering with counterfactual contrastive training," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15965. Springer Nature Switzerland, 2025, pp. 194–204. [Online]. Available: https://doi.org/10.1007/978-3-032-04978-0_19

[15] H. Zhu, Y. Liu, C. Zhou, G. Lu, and B. Chen, "Med-BiasX: Robust medical visual question answering with language biases," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15973. Springer Nature Switzerland, 2025, pp. 369–378. [Online]. Available: https://doi.org/10.1007/978-3-032-05185-1_36

[16] S. Jiang, Y. Chen, S. Song, Y. Zhang, Y. Jin, Y. Feng, J. Wu, and Z. Liu, "Knowing or guessing? robust medical visual question answering via joint consistency and contrastive learning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, ser. Lecture Notes in Computer Science, vol. 15970. Springer Nature Switzerland, 2025, pp. 325–335. [Online]. Available: https://doi.org/10.1007/978-3-032-05141-7_32

[17] L. Butsanets, C. Corbière, J. Khlaut, P. Manceron, and C. Dancette, "RadImageNet-VQA: A large-scale ct and mri dataset for radiologic visual question answering," in *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, vol. 315. PMLR, 2026, pp. 3036–3068.

[18] J. Guasch-Martí, E. Lopez-Cuena, M. Suárez-Fernández, J. Bayarri-Planas, A. Arias-Duart, and D. Garcia-Gasulla, "Aloe-vision: Robust vision-language models for healthcare," in *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, vol. 315. PMLR, 2026, pp. 2404–2426.

[19] B. Gundersen, N. Deperrois, S. Ruiperez-Campillo, T. M. Sutter, J. E. Vogt, M. Moor, F. Nooralahzadeh, and M. Krauthammer, "RadVLM-GRPO: Enhancing chest x-ray report generation and visual grounding via reinforcement learning," in *Proceedings of the 9th International*

Conference on Medical Imaging with Deep Learning, ser. Proceedings of Machine Learning Research, vol. 315. PMLR, 2026, pp. 1935–1968.

[20] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 10 088–10 115.