ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 825-837
DOI 10.5755/j01.itc.55.2.44202

Difference Visual Question Answering in Medical Imaging Based on Multimodal Large Models and Enhanced by Prior Region-level Difference Descriptions

Received 2026/01/31 | Accepted after revision 2026/04/09

HOW TO CITE: Liu, J., Yang, B., Li, H., Huang, H., Li, Y., Cai, Y., Qin, Y. (2026) Difference Visual Question Answering in Medical Imaging Based on Multimodal Large Models and Enhanced by Prior Region-level Difference Descriptions. *Information Technology and Control*, 55(2), 825-837. https://doi.org/10.5755/j01.itc.55.2.44202

Difference Visual Question Answering in Medical Imaging Based on Multimodal Large Models and Enhanced by Prior Region-level Difference Descriptions

Bokai Yang

Institute of Intelligence Science and Engineering, Shenzhen Polytechnic University, Shenzhen 518055, China; e-mail: bk.yang@siat.ac.cn (B.Y.)

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;

Haorong Li

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; e-mail: hr.li@siat.ac.cn (H.L.)

Yirong Qin

School of Urban Governance and Public Affairs, Suzhou City University, Suzhou 215000, China; e-mail: qinyr@szcu.edu.cn (Y.Q.)

Huazhen Huang, Ye Li

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; e-mails: hz.huang@siat.ac.cn (H.H.); ye.li@siat.ac.cn (Y.L.)

Jikui Liu*

Institute of Intelligence Science and Engineering, Shenzhen Polytechnic University, Shenzhen 518055, China; e-mail: liujikui007@gmail.com (J.L.)

Yunpeng Cai*

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; e-mail: yp.cai@siat.ac.cn (Y.C.)

Corresponding author: liujikui007@gmail.com (J. L.); yp.cai@siat.ac.cn (Y. C.)

Difference Visual Question Answering (Diff-VQA) in medical imaging automatically compares patient images across time points to support assessment of lesion progression and treatment efficacy. However, pixel-level matching is unreliable due to non-rigid deformations, view shifts, and acquisition noise, while existing models often rely on synthetic labels and lack effective integration of local and global information. To address these challenges, we propose a multimodal large-model framework that adopts a progressive "local semantic modeling–global difference reasoning" strategy. Key anatomical regions in chest X-rays are localized via object detection and aligned with VinDr-CXR annotations to construct region–disease mappings, transforming misalignment into semantic difference analysis. A dynamic sampling strategy further generates clinically meaningful image pairs with fine-grained difference labels. Finally, a multimodal large model fuses local features with global context to support single-image QA, dual-image disease description, and global difference reasoning. Experiments on the MIMIC-Diff-VQA dataset demonstrate state-of-the-art accuracy in single-image QA and substantial improvements in Diff-VQA tasks over mainstream medical large models. In the single-image QA tasks, our model improves accuracy from 52.5% to 64.1% (22.2% relative improvement), and in the Diff-VQA tasks, the CIDEr score increases from 1.027 to 1.379 (34.3% relative improvement). These results highlight the framework's potential to enhance diagnostic accuracy and strengthen clinical decision support in radiology practice.

KEYWORDS: Large Language Models (LLMs); Multimodal Learning; Visual Question Answering (VQA); Medical Imaging; Difference Visual Question Answering (Diff-VQA); Radiology Report Generation

1. Introduction

Longitudinal comparison of chest radiographs plays a pivotal role in radiology, where clinicians routinely assess disease progression, treatment response, or recurrence by contrasting current and prior examinations [20, 15, 26, 14]. Automated Difference Visual Question Answering (Diff-VQA) offers a promising avenue to support this process by generating interpretable descriptions of temporal changes [31, 32, 27, 6]. Unlike general VQA tasks, Diff-VQA requires precise localization and characterization of disease dynamics, which is particularly valuable in clinical scenarios such as monitoring pneumonia resolution, evaluating tumor growth, or tracking chronic lung conditions [25, 17, 21]. Accurate automation of these comparisons can significantly reduce radiologists' workload while improving diagnostic consistency and efficiency [3, 18].

Early Diff-VQA approaches in the computer vision domain—such as MCCFormers [28], IDCPCL [35], CLIP4IDC [7] and VIXEN [2]—demonstrated the feasibility of modeling visual changes by combining image–text pre-training with difference-aware architectures. However, these methods are optimized for natural images and rely heavily on synthetic editing datasets, which cannot capture the complex anatomical variations or subtle pathological changes encountered in clinical imaging [1, 24, 16]. More importantly, pixel-level feature matching is highly sensitive to non-rigid deformations, patient posture, and acquisition settings, leading to unreliable difference detection in radiology practice [11, 30, 12, 3, 4, 13]. As a result, directly applying these general frameworks to medical imaging yields limited performance [34]. While existing works commonly rely on language-based evaluation metrics such as BLEU, METEOR, ROUGE, and CIDEr to assess the quality of generated difference descriptions, these metrics primarily measure surface-level lexical similarity and may not fully reflect the clinical correctness or semantic coherence of the findings. In medical imaging scenarios, diagnostically meaningful interpretations can often be expressed using diverse wording, and a higher degree of lexical overlap does not necessarily imply accurate clinical reasoning. Therefore, these language-only metrics should be interpreted with caution when evaluating Diff-VQA systems, and performance improvements reported under such metrics should be contextualized with an understanding of their inherent limitations.

To address this gap, recent medical-specific Diff-VQA frameworks have attempted to incorporate anatomical priors and knowledge graphs [5, 9, 29,

19]. While they improve global reasoning, existing methods often suffer from redundant region definitions, insufficient sensitivity to long-range anatomical dependencies, and high computational cost [23, 8, 22]. Furthermore, most approaches do not adequately integrate local lesion-level descriptions with global reasoning, resulting in limited clinical interpretability.

In this study, we propose a multimodal large-model framework for medical Diff-VQA enhanced by prior region-level difference descriptions [33]. Our key insight is a progressive "local semantic modeling–global difference reasoning" strategy: anatomical regions are first localized and associated with disease annotations, transforming image misalignment into high-level semantic analysis; these structured regional descriptions then guide global reasoning, enabling more accurate, interpretable, and clinically coherent results. By bridging lesion-level findings and overall disease trajectories, the framework better aligns with the diagnostic workflow of radiologists and provides meaningful decision support [10].

The main contributions of this work are threefold:

We introduce a staged learning mechanism that combines local semantic modeling with global difference reasoning, mitigating the impact of acquisition variability while enhancing temporal comparison.

We design a dynamic image-pair sampling and fine-grained labeling strategy to generate clinically meaningful difference supervision.

We validate our approach on the large-scale MIMIC-Diff-VQA dataset, demonstrating superior performance over existing methods in both single-image VQA and multi-image Diff-VQA tasks, with improved interpretability and clinical value.

2. Methodology

2.1. Material

To train and evaluate the proposed medical image Diff-VQA framework, we utilized three publicly available large-scale chest radiograph datasets: MIMIC-Diff-VQA, Chest ImaGenome, and VinDr-CXR, as shown in Table 1.

The MIMIC-Diff-VQA dataset is a publicly available, deidentified subset of the MIMIC-CXR database released via the PhysioNet platform. The MIMIC-CXR dataset and all its derivatives have received IRB approval from the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center, ensuring compliance with HIPAA regulations. The MIMIC-Diff-VQA dataset comprises 164,324 image pairs, where each pair includes a main image and a reference image. It contains a total of 700,703 curated question–answer (QA) pairs categorized into seven types: abnormality (145,421), presence (155,726), view (56,265), location (84,193), type (27,478), level (67,296), and difference (164,324). The first six categories correspond to single-image VQA tasks, where each QA pair is associated with one image (either the main or the reference), resulting in 482,698 training samples and 53,681 test samples. The last category, difference, defines multi-image Diff-VQA tasks, in which each QA pair involves both the main and reference images of an image pair, yielding 131,462 training pairs and 32,862 test pairs. All QA pairs were derived from radiology reports and verified by clinical experts to ensure medical accuracy and contextual consistency. The dataset is specifically designed to support both single-image reasoning and cross-image temporal comparison, making it ideally suited for developing and evaluating medical Diff-VQA models.

The Chest ImaGenome dataset provides structured annotations for over 240,000 chest X-ray images. Each image is annotated with 29 anatomical locations and associated attributes, including location, texture, and pathological findings. Moreover, the dataset includes over 670,000 comparative relations to represent changes in anatomical features across different time points. The annotations were generated through a combination of natural language processing (NLP) techniques and image-based segmentation methods. This dataset is particularly useful for region-level semantic modeling, enabling the construction of scene graphs that reflect complex anatomical structures and their temporal evolution.

VinDr-CXR is a large-scale chest X-ray dataset designed to facilitate research on thoracic disease detection and localization. It contains over 100,000 frontal chest X-ray scans, among which 18,000 images were manually annotated by 17 board-certified radiologists. The dataset provides both global labels (e.g., pneumonia, tuberculosis) and local bounding-box annotations that pinpoint disease-specific

Table 1
Summary of datasets.

Dataset	Data Volume	Annotation Types	Key Features	Application in This Study
MIMIC-Diff-VQA	164,324 images; >700,000 QA pairs	Seven question categories	Derived from MIMIC-CXR; expert-verified QA pairs	Training and evaluation for temporal reasoning
Chest ImaGenome	>240,000 images	29 anatomical locations; >670,000 relations	Structured region-level annotations	Building region-level semantic model
VinDr-CXR	>100,000 images (18,000 manually annotated)	Global disease labels plus local bounding boxes	High-fidelity labels, expert consensus on test set	Region–disease mapping and fine-grained supervision

regions of interest. The training and testing splits include 15,000 and 3,000 images, respectively. Each training image is labeled independently by three radiologists, while test images are annotated through expert consensus, ensuring high labeling fidelity. This dataset supports the learning of fine-grained pathology localization and is instrumental in constructing disease-region mappings for semantic VQA.

We strictly enforce subject-level (patient-level) splits across all stages—including de-tector training (Chest ImaGenome/VinDr-CXR), local difference construction (pair sam-pling), LVLM fine-tuning (MIMIC-Diff-VQA), and final evaluation—so that no patient ap-pears in more than one split. For longitudinal exams, all time points from the same subject are assigned to the same split, thereby eliminating temporal overlap or leakage across train and test sets.

2.2. Methods

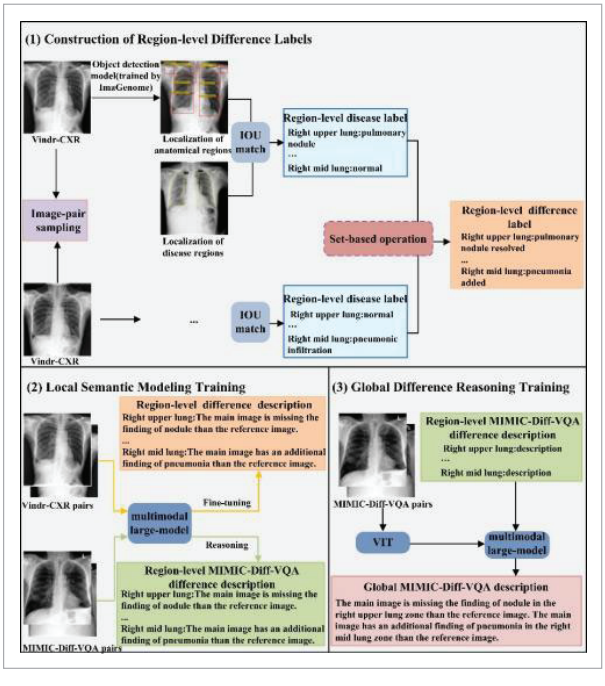
The objective of this study is to address VQA and Diff-VQA tasks in medical imaging and to propose a multimodal large-model training and inference framework based on prior region-level difference description. The framework is designed to handle single-image multi-type VQA and multi-image Diff-VQA within a unified system. As illustrated in Figure 1, our approach comprises three major components: (i) construction of region-level difference labels, (ii) local semantic modeling training, and (iii) global difference reasoning training.

2.2.1. Construction of Region-level Difference Labels

First, an object detection model was trained using the anatomical region annotations from the Chest ImaGenome dataset. This detector identifies eight predefined anatomical regions—right upper lung zone, right mid lung zone, right lower lung zone, left upper lung zone, left mid lung zone, left lower lung zone, right clavicle, and left clavicle—while other regions are grouped as "Others." These eight regions were selected to ensure broad coverage of the thoracic cavity and to minimize redundancy, thereby providing adequate spatial resolution for downstream disease diagnosis and difference analysis.

Second, to establish a region–disease correspondence, the detection results from the trained model were aligned with disease annotations from the VinDr-CXR dataset. Through spatial overlap analysis, a region–disease mapping was built. Disease bounding

Figure 1
Overview of Our Proposed Method.



boxes are assigned to anatomical regions based on spatial overlap, using an IoU threshold of 0.3. Specifically, a disease is assigned to the region whose bounding box has IoU ≥ 0.3 with the disease bounding box. In cases where multiple regions satisfy the threshold, the disease is assigned to the region with the maximum IoU value. If no region meets the threshold, the disease is labeled as "Others" and does not contribute to region-level difference computation. This process produced, for each chest X-ray image, a structured table documenting the association between disease labels and specific anatomical regions.

Third, to construct high-quality Diff-VQA training pairs, we developed an image-pair sampling strategy that jointly optimizes visual similarity and disease-label divergence. The objective is to identify clinically meaningful image pairs that reflect pathological changes while maintaining overall structural consistency, enabling the model to learn clinically relevant differences rather than irrelevant variations. Concretely, for each image pair $(I_i,I_j)(I_i,I_j)$, composite score S_{ij} is defined as:

$$S_{ij} = \alpha \cdot Sim(I_i, I_j) + (1-\alpha) \cdot Div(L_i, L_j) \quad (1)$$

where α balances the contribution of visual similarity and disease-label divergence. The similarity term $Sim(I_i,I_j)$ is computed by extracting features from each image using a pre-trained ResNet-50 and then calculating the cosine similarity between feature vectors. The divergence term $Div(L_i,L_j)$ is measured by the Jaccard distance between the disease-label sets L_i and L_j:

$$Div(L_i, L_j) = 1 - \frac{|A \cap B|}{|A \cup B|}, \quad (2)$$

where A and B denote the sets of disease labels for two images. A higher Jaccard distance indicates greater divergence, capturing more substantial pathological differences and thus offering more informative comparisons for clinical training.

Finally, region-level difference labels were generated to provide fine-grained supervision. For each anatomical region, the disease-label sets from paired images (I_i,I_j) were compared using set operations to derive newly appeared, resolved, or persistent findings:

$$D_{add} = D_j - (D_i \cap D_j) \quad (3)$$

$$D_{remove} = D_i - (D_i \cap D_j) \quad (4)$$

$$D_{keep} = D_i \cap D_j \ , \quad (5)$$

where D_{add} (New disease set) represents diseases present in image j but absent in image T; it captures new pathological findings appearing between two time points. D_{remove} (Resolved disease set) represents diseases present in image i but absent in image j; it captures previously existing pathologies that have resolved or disappeared. D_{keep} (Persistent disease set) represents diseases present in both images; it captures stable or persistent pathological findings across two time points.

This structured representation allows fine-grained tracking of disease dynamics—appearance, resolution, and persistence—at the anatomical-region level, providing rich supervision for model training and clinically interpretable outputs.

2.2.2. Local Semantic Modeling Training

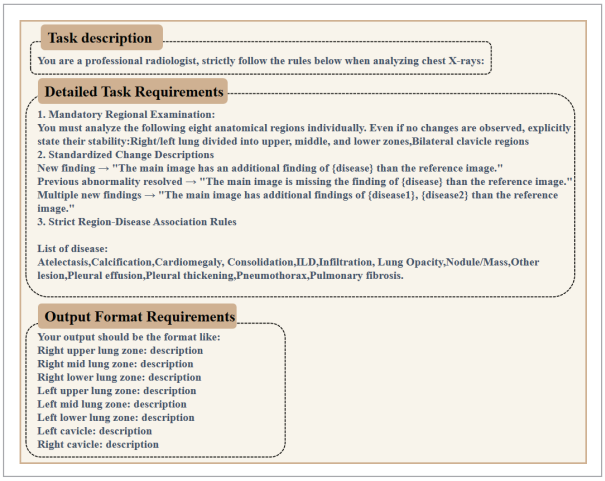
This stage guides the large-model to generate region-level difference description by comparing temporally paired chest radiographs. Using a prompt-based supervision strategy (Figure 2), the model is explicitly informed which image is the "primary" (current) image and which is the "reference" image, ensuring correct temporal interpretation. The prompt instructs the model to act as a professional radiologist and sequentially examine the eight predefined anatomical regions (upper, middle, and lower zones of both lungs, plus bilateral clavicular regions). Even when no change is observed, the model must explicitly state the region's stability to guarantee completeness.

Difference descriptions follow a standardized template. New abnormalities are expressed as "Compared with the reference image, the current image shows a new finding of {disease}." Resolved abnormalities are expressed as "The current image lacks the previously observed {disease} from the reference image." Multiple abnormalities are concatenated in order. To enhance accuracy, the prompt includes a predefined disease list and requires all outputs to

use a structured template that explicitly links each anatomical region with its disease changes. This design enforces precise region–disease alignment and prevents redundant or ambiguous outputs when processing the same image pair.

Figure 2

Local Prompt Design for Region-level Difference Description.



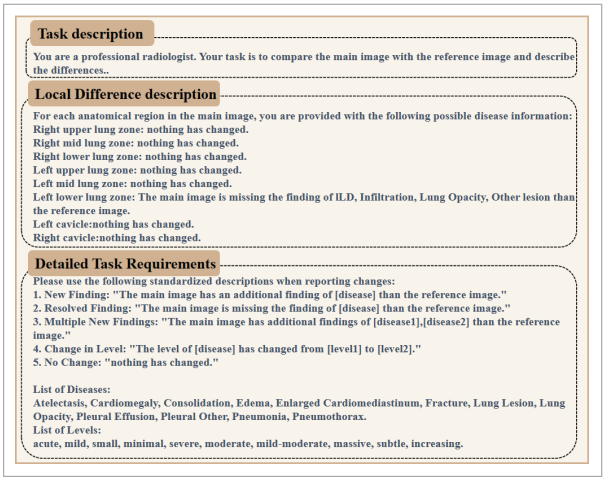
2.2.3. Global Difference Reasoning Training

In this stage, the model performs comprehensive image-level comparisons by integrating region-level difference description as prior knowledge into global reasoning. Similar to the local setup, the prompt explicitly specifies the "primary" and "reference" images to establish temporal directionality. As shown in Figure 3, the model is instructed to act as a professional radiologist and produce an overall comparative report describing differences between the primary and reference images.

For each anatomical region, the model incorporates precomputed local difference description and contextual disease information to perform higher-level semantic reasoning. The prompt employs a structured template containing a predefined disease vocabulary and severity levels (e.g., mild, moderate, severe, unchanged). These additional attributes enable the model to detect not only binary changes (appearance/disappearance) but also gradations of severity and progression. By combining regional findings with severity assessment, the model learns to generate clinically coherent global diagnostic summaries that reflect both spatial and pathological dynamics.

Figure 3

Global Prompt Design for Cross-Image Difference Description.



3. Results and Discussion

To ensure fair comparison, all baseline models were trained and evaluated using the same subject-level train/validation/test splits as our method. For both single-image and temporal difference VQA tasks, we employed a unified prompt template and kept the question formatting identical across all models. All baselines were fine-tuned on the same training samples and evaluated on the same held-out subject-level test set. During inference, we used a deterministic decoding configuration to eliminate variability across models.

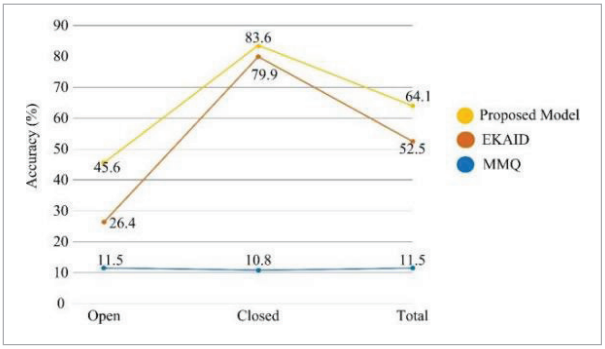
3.1. Performance of Single-Image VQA Tasks

In the single-image VQA task, the dataset comprises 482,698 training samples and 53,681 test samples, covering six major question types: abnormality, presence, view, location, type, and level. These categories comprehensively evaluate various aspects of chest X-ray interpretation. Specifically, abnormality questions assess whether any pathological signs are present; presence questions focus on detecting disease in specific anatomical regions; view questions examine the radiographic projection, which is crucial for diagnostic accuracy due to the influence of imaging angles on visual details; location questions aim to identify the precise position of abnormalities; type questions target the classification of disease categories; and level questions evaluate the severity of dis-

ease, contributing to progression monitoring. Within this framework, questions requiring binary responses (e.g., "yes" or "no") are defined as closed-ended, such as "Is there disease in the left upper lung zone?", while all others demanding detailed responses regarding disease type, anatomical location, or severity are categorized as open-ended. As shown in Figure 4, our proposed model achieved the highest accuracy across all categories on the test set, demonstrating superior performance in both classification and descriptive tasks within the single-image VQA setting.

Figure 4

Performance comparison of models on single-image VQA Tasks.



3.2. Performance of Multi-Image Diff-VQA Tasks

In the MIMIC-Diff-VQA dataset, multi-image Diff-VQA tasks are referred to as "difference" questions, which center on comparing a patient's current and prior chest X-ray images to identify differences in disease manifestations. These questions are designed to analyze whether new pathologies have appeared, existing findings have resolved, or previously detected conditions have progressed or regressed in terms of severity or grade. Addressing such tasks requires not only precise visual comparison of image details, but also the application of domain-specific medical knowledge to classify pathological changes and assess them quantitatively. This setup reflects realistic clinical scenarios, where radiologists routinely compare historical and follow-up images to monitor disease progression or evaluate therapeutic response. As such, Diff-VQA provides valuable support for clinical diagnosis, treatment evaluation, and follow-up planning by offering a structured and interpretable summary of a patient's disease trajectory.

Table 2 presents a performance comparison among various models on the Diff-VQA task using the MIMIC-Diff-VQA test set. LLaVA and Med-Flamingo represent general-purpose vision-language models and domain-adapted medical models, respectively. In contrast, MultiMedRes (which employs large models as expert agents for knowledge aggregation and reasoning) and EKAID (which leverages graph neural networks for disease-region representation) represent current state-of-the-art approaches tailored for medical image Diff-VQA. As shown in the results, our proposed method significantly outperforms all baselines across multiple evaluation metrics. Our method addresses these limitations by optimizing region selection to improve the quality of prior knowledge and reduce irrelevant influence. Furthermore, by enhancing the model's ability to utilize spatial informa-

Table 2

Performance comparison of models on multi-image Diff-VQA Tasks (difference question).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDEr
MCCFormers	0.214	0.190	0.170	0.153	0.319	0.340	0
LLaVA	0.411	0.333	0.257	0.185	0.320	0.452	0.162
Med-flamingo	0.573	0.505	0.433	0.359	0.304	0.544	0.438
IDCPCL	0.614	0.541	0.474	0.414	0.303	0.582	0.703
MultiMedRes	0.610	0.535	0.473	0.418	0.357	0.586	0.843
EKAID	0.628	0.553	0.491	0.434	0.339	0.577	1.027
Ours	0.661	0.584	0.523	0.467	0.379	0.625	1.379

Table 3
Performance comparison of models on single-image VQA Tasks and multi-image Diff-VQA Tasks (all question).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDEr
EKAID	0.626	0.541	0.477	0.422	0.340	0.645	1.911
Coarse-to-Fine	0.630	0.543	0.479	0.422	0.403	0.662	2.022
Ours	0.653	0.562	0.488	0.435	0.399	0.675	2.242

tion, we implement a progressive difference reasoning strategy from local to global levels, enabling the model to better capture subtle inter-image contrasts. As a result, our approach achieves the best overall performance on the Diff-VQA task.

Table 3 presents the performance comparison of different models on the entire set of question types in the MIMIC-Diff-VQA test dataset. As shown, our proposed method achieves significant improvements over EKAID across all evaluation metrics. Compared with the Coarse-to-Fine approach, our method also demonstrates clear advantages on most indicators. These results indicate that the proposed framework offers superior generalization ability and robustness when handling a wide variety of visual question answering tasks in the medical imaging domain.

To assess the clinical reliability of the generated difference statements, we evaluate whether the model correctly identifies "new", "resolved", and "keep" findings for each anatomical region.

Using the constructed region-level difference labels (D_{add}, D_{remove}, D_{keep}) as ground truth, we compare the predicted change type for each region and report precision, recall, and F1-score. As shown in Table 4, the proposed method achieves an average F1-score of 0.85, which is significantly higher than EKAID.

This demonstrates that our framework accurately captures lesion evolution and is clinically meaningful beyond textual similarity metrics.

To further illustrate how the model performs in real clinical scenarios, we provide qualitative examples that compare the ground truth and our proposed model under both global difference descriptions and region-level local reasoning. These examples include (i) typical correct cases, where the proposed model accurately identifies clinically meaningful changes, and (ii) failure cases (false positives), where the model over-interprets image variations. Presenting both success and failure cases allows us to examine how the model's predictions align with radiological judgment and to highlight potential clinical risks, particularly when detecting subtle parenchymal or morphological differences. Figure 5 and Figure 6 illustrate these behaviors at both the global summary level and the fine-grained anatomical region level.

Figure 5
Qualitative examples of global difference descriptions.

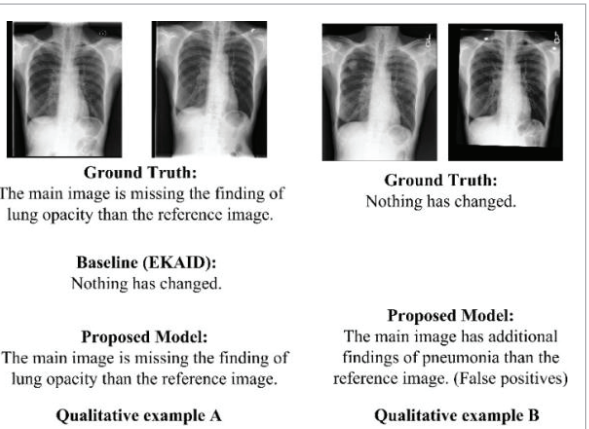
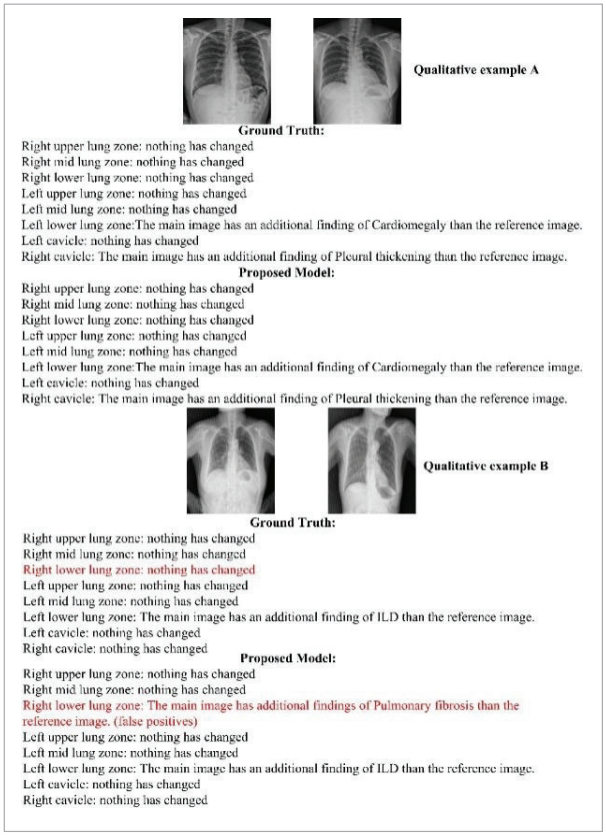


Table 4
Region-level clinical accuracy of change-type prediction.

Method	F1-score	Precision	Recall	P-value vs EKAID (F1)
EKAID	0.76	0.78	0.74	-
Ours	0.85	0.88	0.83	<0.001

Figure 6
Qualitative examples of region-level difference descriptions.



3.3. Ablation Study

To further investigate the contribution of individual components in our proposed framework, we conducted a series of ablation experiments based on the MIMIC-Diff-VQA dataset, focusing on the Diff-VQA task (Table 4). The goal is to assess how the integration of prior spatial knowledge and our refined regional reasoning strategy affects overall model performance.

The Baseline model refers to a standard configuration in which a large-scale multimodal model is fine-tuned on the Diff-VQA dataset without incorporating any anatomical priors or spatial constraints. This setting provides a reference point to understand the inherent capability of the pretrained model in performing difference reasoning purely based on image-text inputs. Building on this, the Base-Graph variant integrates spatial prior information derived from EKAID's disease-region graph module. This graph encodes predefined anatomical regions and their relationships to guide the model in identifying spatially relevant disease patterns. Experimental results show that Base-Graph achieves notable performance improvements over the baseline, demonstrating the effectiveness of introducing structural spatial knowledge. However, we observed that graph-based priors are susceptible to redundant region mappings and interference from irrelevant negative samples, especially in complex medical scenarios where lesions span multiple anatomical boundaries or overlap with normal structures. To address this issue, our full model introduces an optimized region selection mechanism that reduces noise from irrelevant areas and enhances the precision of spatial supervision. Furthermore, by designing a progressive reasoning framework from local to global comparison levels, our method enables more accurate and interpretable multi-image difference analysis.

The ablation results highlight the necessity of incorporating both cleaned anatomical priors and hierarchical reasoning strategies to fully leverage the capabilities of large multimodal models in complex medical QA tasks, see Table 5.

To further analyze the impact of anatomical region granularity on our framework, we conducted a set of ablation experiments on the MIMIC-Diff-VQA dataset, focusing on the Diff-VQA task with varying numbers of regions (K=4, 6, 8, 10, 16, 28). The goal is to determine how the choice of K influences the balance between fine-grained localization and noise propagation in regional reasoning.

Table 5
Ablation study of models on multi-image Diff-VQA Tasks.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDEr
Baseline	0.649	0.573	0.512	0.455	0.366	0.609	1.206
Base-Graph	0.653	0.580	0.521	0.462	0.371	0.618	1.267
Ours	0.661	0.584	0.523	0.467	0.379	0.625	1.379

Table 6
Ablation study on the number of anatomical regions (K).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDEr
K=4	0.635	0.562	0.498	0.438	0.370	0.610	1.165
K=6	0.648	0.574	0.507	0.447	0.375	0.620	1.210
Ours (K=8)	0.661	0.584	0.523	0.467	0.379	0.625	1.379
K=10	0.655	0.578	0.515	0.455	0.378	0.617	1.330
K=16	0.645	0.565	0.495	0.425	0.373	0.605	1.260
K=28	0.628	0.550	0.475	0.400	0.365	0.590	1.190

Table 7
Ablation study of the weighting parameter α in the similarity-divergence sampling strategy.

Method	BLEU	METEOR	ROUGE_L	CiDEr	P-value vs α =0.5 (CIDEr)
α=0.3	0.447	0.375	0.620	1.210	<0.001
Ours (α=0.7)	0.467	0.379	0.625	1.379	-
α=0.7	0.455	0.378	0.617	1.330	0.039

In this series of experiments, we kept all other configurations identical to the default setting and only varied the number of anatomical regions. Results in Table 6 show a clear trend: performance improves consistently from K=4 to K=6 and reaches its peak at K=8, where the model achieves the highest scores across BLEU, METEOR, ROUGE-L, and CIDEr. However, when further increasing K beyond 8, the results begin to decline, with noticeable drops at K=10, K=16, and K=28. This suggests that while moderate granularity (K=8) provides sufficient anatomical detail to capture subtle regional differences, excessive subdivision introduces redundant regions and amplifies localization errors. Such noise propagates through the region-disease mapping and ultimately degrades text generation quality in the Diff-VQA task.

These findings confirm that K=8 strikes the optimal balance between representation power and robustness, maximizing the benefit of anatomical priors while avoiding over-segmentation pitfalls. The ablation study therefore highlights the importance of carefully selecting anatomical granularity in order to fully leverage large multimodal models for accurate and interpretable medical difference question answering.

To evaluate the influence of the weighting parameter α in the image-pair sampling strategy, we conducted an ablation experiment on the MIMIC-Diff-VQA dataset, focusing on the Diff-VQA task with varying $\alpha \in \{0.3, 0.5, 0.7\}$. α controls the trade-off between the Sim and the Div when selecting clinically meaningful image pairs. A higher α emphasizes visual similarity, while a lower α emphasizes pathological difference diversity.

Table 7 reports the performance under different α settings. In this series of experiments, we kept all other configurations identical to the default setting and only varied α. The results show that our framework is stable across a wide range of values, indicating that the sampling strategy is robust. The best overall performance is achieved at α = 0.5, which balances structural consistency with clinically relevant disease progression. Therefore, α = 0.5 is used as the default configuration in all main experiments.

In addition, we performed an ablation study on the backbone used for computing visual similarity (Sim) within the pair-selection score S_{ij}. Besides the default ResNet-50, we evaluated Swin-Transformer and ViT-Base backbones to assess the sensitivity of the similarity representation. All backbones were pretrained on ImageNet and fine-tuned under identical training conditions. As shown in Table 8, the performance differences across backbones are within ±1.8%, demonstrating that the proposed framework is robust to the choice of feature extractor for

Table 8
Ablation study on the backbone used for computing visual similarity in the pair-selection score S_{ij}.

Method	BLEU	METEOR	ROUGE_L	CiDEr	P-value vs ResNet-50 (CIDEr)
ViT	0.462	0.377	0.621	1.355	0.11
Swin	0.465	0.378	0.623	1.372	0.47
Ours (ResNet-50)	0.467	0.379	0.625	1.379	-

similarity computation. ResNet-50 is selected as the default backbone for its favorable trade-off between accuracy and efficiency.

4. Conclusion

This study proposes a multimodal large-model framework for medical Diff-VQA, enhanced by prior region-level difference descriptions. Instead of relying on fragile pixel-level comparisons, the framework elevates longitudinal image comparison to high-level semantic reasoning, thereby better handling non-rigid deformations, view shifts, and imaging variability commonly seen in clinical chest radiographs. With a progressive "local semantic modeling–global difference reasoning" strategy, the model first grounds changes at anatomically localized regions and links them to disease mappings, then aggregates these cues into global temporal reasoning. This design enables both fine-grained interpretability and robust diagnostic inference. Extensive experiments on the MIMIC-Diff-VQA dataset show that the proposed method consistently outperforms existing state-of-the-art approaches in both single-image VQA and dual-image Diff-VQA tasks. Ablation studies further validate the contributions of region-level priors and staged reasoning. Importantly, beyond improving quantitative performance, the framework produces interpretable region-level descriptions of disease changes, supporting clinical understanding and verification.

From a clinical perspective, the framework can assist radiologists in longitudinal comparison by accelerating detection of newly emerged or resolved findings, reducing interpretation variability, and enabling more consistent disease monitoring. By providing structured regional change descriptions together with global comparative summaries, the system enhances transparency and facilitates human–AI collaboration in diagnostic workflows—an essential step toward building clinical trust and practical deployment. Nevertheless, limitations remain: real-world performance still requires prospective multi-center validation. We are currently planning a staged prospective evaluation in collaboration with two regional medical centers. Phase I (0–6 months) will focus on retrospective multi-center validation with harmonized labeling protocols. Phase II (6–18 months) will involve prospective data collection and real-time clinical decision support trials under clinician supervision. These steps will allow us to characterize the model's robustness, inter-hospital transferability, and workflow integration feasibility; reliance on existing annotations may miss subtle or overlapping pathologies; and future work should incorporate additional multimodal signals such as electronic medical records, laboratory results, and physician notes to achieve more holistic clinical reasoning. Overall, this work advances medical Diff-VQA and serves as a proof-of-concept toward future integration into radiology workflows, supporting efficient and accurate assessment of disease progression or remission.

Acknowledgements

The authors acknowledge Shenzhen Science and Technology Program (No. KJZD20240903102802003), the Post-doctoral Foundation Project of Shenzhen Polytechnic University (No.6024331020K), open research fund of Suzhou Key Lab of Multi-modal Data Fusion and Intelligent Healthcare (25SZZD08), the Shenzhen Polytechnic University Research Fund (No. 6023312038K), the Special projects fund in key areas of the Guangdong Provincial Department of Education (2023ZDZX2085), Shenzhen major science and technology project (KCXFZ20240903093923031) and the National Natural Science Foundation of China under grant (No. U22A2041).

References

1. Al-Hadhrami, S., Menai, M. E. B., Al-Ahmadi, S., Alnafessah, A. A Critical Analysis of Benchmarks, Techniques, and Models in Medical Visual Question Answering. IEEE Access, 2023, 11: 136507-136540. https://doi.org/10.1109/ACCESS.2023.3335216
2. Black, A., Shi, J., Fan, Y., Bui, T., Collomosse, J. Vixen: Visual Text Comparison Network for Image Difference Captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(2): 846-854. https://doi.org/10.1609/aaai.v38i2.27843
3. Chizhikova, M., López-Úbeda, P., Martín-Noguerol, T., Díaz-Galiano, M. C., Ureña-López, L. A., Luna, A., Martín-Valdivia, M. T. Automatic TNM Staging of Colorectal Cancer Radiology Reports Using Pre-Trained Language Models. Computer Methods and Programs in Biomedicine, 2025, 259: 108515. https://doi.org/10.1016/j.cmpb.2024.108515
4. Darzi, F., Bocklitz, T. A Review of Medical Image Registration for Different Modalities. Bioengineering, 2024, 11(8): 786. https://doi.org/10.3390/bioengineering11080786
5. Gu, Z., Liu, F., Yin, C., Zhang, P. Inquire, Interact, and Integrate: A Proactive Agent Collaborative Framework for Zero-Shot Multimodal Medical Reasoning. arXiv Preprint arXiv:2405.11640, 2024. https://doi.org/10.1002/aisy.202400840
6. Guo, E., Zhao, Z., Wang, Z., Chen, T., Liu, Y., Zhou, L. DiN: Diffusion Model for Robust Medical VQA with Semantic Noisy Labels. arXiv Preprint arXiv:2503.18536. https://doi.org/10.48550/arXiv.2503.18536
7. Guo, Z., Wang, T. J., Laaksonen, J. CLIP4IDC: CLIP for Image Difference Captioning. arXiv Preprint arXiv:2206.00629, 2022. https://doi.org/10.18653/v1/2022.aacl-short.5
8. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv Preprint arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685
9. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R. M., Zhu, Y. Expert Knowledge-Aware Image Difference Graph Representation Learning for Difference-Aware Medical Visual Question Answering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 4156-4165. https://doi.org/10.1145/3580305.3599819
10. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R. M., Zhu, Y. Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images. PhysioNet, 2023, 12: 13. https://doi.org/10.13026/e6dd-cn74
11. Hua, R. Non-Rigid Medical Image Registration with Extended Free Form Deformations: Modelling General Tissue Transitions. University of Sheffield, 2016. https://etheses.whiterose.ac.uk/id/eprint/17710/
12. Im, J. E., Khalifa, M., Gregory, A. V., Erickson, B. J., Kline, T. L. A Systematic Review on the Use of Registration-Based Change Tracking Methods in Longitudinal Radiological Images. Journal of Imaging Informatics in Medicine, 2024: 1-14. https://doi.org/10.1007/s10278-024-01333-1
13. Iwao, Y., Kawata, N., Sekiguchi, Y., Haneishi, H. Non-rigid Registration Method for Longitudinal Chest CT Images in COVID-19. Heliyon, 2024, 10(17). https://doi.org/10.1016/j.heliyon.2024.e37272
14. Jantscher, M., Gunzer, F., Reishofer, G., Kern, R. Causal Insights from Clinical Information in Radiology: Enhancing Future Multimodal AI Development. Computer Methods and Programs in Biomedicine, 2025: 108810. https://doi.org/10.1016/j.cmpb.2025.108810
15. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Mark, R. G., Horng, S. MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports. Scientific Data, 2019, 6(1): 317. https://doi.org/10.1038/s41597-019-0322-0
16. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2901-2910. https://doi.org/10.1109/CVPR.2017.215
17. Katsuragawa, S., Doi, K. Computer-Aided Diagnosis in Chest Radiography. Computerized Medical Imaging and Graphics, 2007, 31(4-5): 212-223. https://doi.org/10.1016/j.compmedimag.2007.02.003
18. Li, J., Zhou, T., Zhou, Z., Wei, X., Song, H., Wang, Z., Chen, Y., Lv, H. Experience-Guided Multi-Agent Interpretable Framework for Radiology Report Summarization. Computer Methods and Programs in Biomedicine, 2025: 109078. https://doi.org/10.1016/j.cmpb.2025.109078

19. Liang, X., Wang, Y., Wang, D., Jiao, Z., Zhong, H., Yang, M., Wang, Q. Leveraging Coarse-to-Fine Grained Representations in Contrastive Learning for Differential Medical Visual Question Answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024: 415-425. https://doi.org/10.1007/978-3-031-72086-4_39

20. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z. Medical Visual Question Answering: A Survey. Artificial Intelligence in Medicine, 2023, 143: 102611. https://doi.org/10.1016/j.artmed.2023.102611

21. Liu, F., Panagiotakos, D. Real-World Data: A Brief Review of the Methods, Applications, Challenges and Opportunities. BMC Medical Research Methodology, 2022, 22(1): 287. https://doi.org/10.1186/s12874-022-01768-6

22. Mao, Y., Ge, Y., Fan, Y., Xu, W., Mi, Y., Hu, Z., Gao, Y. A Survey on LoRA of Large Language Models. Frontiers of Computer Science, 2025, 19(7): 197605. https://doi.org/10.1007/s11704-024-40663-9

23. Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., Ngo, D. T., Nguyen, N. T., Nguyen, N. T., Dao, M., Vu, V. VinDr-CXR: An Open Dataset of Chest X-Rays with Radiologist's Annotations. Scientific Data, 2022, 9(1): 429. https://doi.org/10.1038/s41597-022-01498-w

24. Park, D. H., Darrell, T., Rohrbach, A. Robust Change Captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4624-4633. https://doi.org/10.1109/ICCV.2019.00472

25. Park, J., Phang, J., Shen, Y., Wu, N., Kim, S., Moy, L., Cho, K., Geras, K. J. Screening Mammogram Classification with Prior Exams. arXiv Preprint arXiv:1907.13057, 2019. https://doi.org/10.48550/arXiv.1907.13057

26. Pham, H. H., Nguyen, N. H., Tran, T. T., Nguyen, T. N., Nguyen, H. Q. PediCXR: An Open, Large-Scale Chest Radiograph Dataset for Interpretation of Common Thoracic Diseases in Children. Scientific Data, 2023, 10(1): 240. https://doi.org/10.1038/s41597-023-02102-5

27. Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., Conant, E. F., Fajardo, L. L., Bassett, L., Dorsi, C., Jong, R., Rebner, M. Diagnostic Performance of Digital Versus Film Mammography for Breast-Cancer Screening. New England Journal of Medicine, 2005, 353(17): 1773-1783. https://doi.org/10.1056/NEJMoa052911

28. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y. Describing and Localizing Multiple Changes with Transformers. arXiv Preprint arXiv:2103.14146. https://doi.org/10.48550/arXiv.2103.14146

29. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

30. Sotiras, A., Davatzikos, C., Paragios, N. Deformable Medical Image Registration: A Survey. IEEE Transactions on Medical Imaging, 2013, 32(7): 1153-1190. https://doi.org/10.1109/TMI.2013.2265603

31. Taylor-Phillips, S., Wallis, M. G., Duncan, A., Gale, A. G. Use of Prior Mammograms in the Transition to Digital Mammography: A Performance and Cost Analysis. European Journal of Radiology, 2012, 81(1): 60-65. https://doi.org/10.1016/j.ejrad.2010.10.025

32. Timberg, P., Hellgren, G., Dustler, M., Tingberg, A. Investigating the Effect of Adding Comparisons with Prior Mammograms to Standalone Digital Breast Tomosynthesis Screening. Journal of Medical Imaging, 2025, 12(S2): S22003-S22003. https://doi.org/10.1117/1.JMI.12.S2.S22003

33. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv Preprint arXiv:2409.12191, 2024. https://doi.org/10.48550/arXiv.2409.12191

34. Xu, R., Jiang, P., Luo, L., Xiao, C., Cross, A., Pan, S., Sun, J., Yang, C. A Survey on Unifying Large Language Models and Knowledge Graphs for Biomedicine and Healthcare. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025: 6195-6205. https://doi.org/10.1145/3711896.3736556

35. Yao, L., Wang, W., Jin, Q. Image Difference Captioning with Pre-Training and Contrastive Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 3108-3116. https://doi.org/10.1609/aaai.v36i3.20218