Spot the Difference: Difference Visual Question Answering with Residual Alignment

Zilin Lu[1*], Yutong Xie[2*], Qingjie Zeng[1], Mengkang Lu[1], Qi Wu[2], and Yong Xia[1,3,4†]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Australian Institute for Machine Learning, The University of Adelaide, Australia
[3] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
[4] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn

Abstract. Difference Visual Question Answering (DiffVQA) introduces a new task aimed at understanding and responding to questions regarding the disparities observed between two images. Unlike traditional medical VQA tasks, DiffVQA closely mirrors the diagnostic procedures of radiologists, who frequently conduct longitudinal comparisons of images taken at different time points for a given patient. This task accentuates the discrepancies between images captured at distinct temporal intervals. To better address the variations, this paper proposes a novel **Re**sidual **Al**ignment model (ReAl) tailored for DiffVQA. ReAl is designed to produce flexible and accurate answers by analyzing the discrepancies in chest X-ray images of the same patient across different time points. Compared to the previous method, ReAl additionally aid a residual input branch, where the residual of two images is fed into this branch. Additionally, a Residual Feature Alignment (RFA) module is introduced to ensure that ReAl effectively captures and learns the disparities between corresponding images. Experimental evaluations conducted on the MIMIC-Diff-VQA dataset demonstrate the superiority of ReAl over previous state-of-the-art methods, consistently achieving better performance. Ablation experiments further validate the effectiveness of the RFA module in enhancing the model's attention to differences. The code implementation of the proposed approach will be made available.

Keywords: Diffenence VQA · Generative model · Residual feature alignment.

[*] Z. Lu and Y. Xie contributed equally to this work.

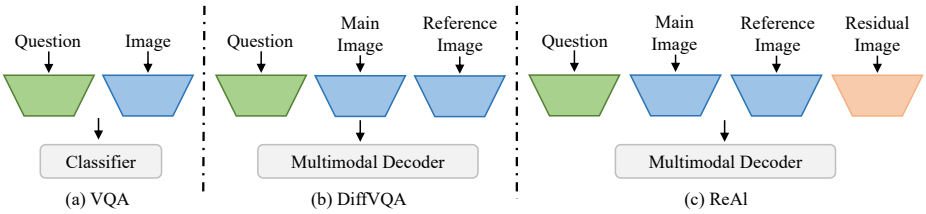

Fig. 1: Comparison of different VQA architectures. All of them have one question input but different numbers of image inputs. Compared to traditional VQA (a), DiffVQA (b) has an additional input called reference image as the standard forjudging difference. Our proposed ReAl (c) calculate the residual of two images as the residual input. Thanks to the residual branch, ReAl is able to focus on differential region.

1 Introduction

Medical imaging plays a pivotal role in modern healthcare for diagnostic and therapeutic purposes [14]. Through various imaging modalities such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), among others, medical imaging provides non-invasive means for physicians to study and assess the internal structures and functions of patients. In clinical diagnosis, radiologists play a crucial role in diagnosing and monitoring various pathologies through the analysis of imaging data collected over time. Their diagnostic process often involves meticulous comparisons of images obtained at different time points to discern subtle changes indicative of disease progression or treatment response. This practice, akin to a longitudinal assessment, underscores the importance of understanding the nuances between image pairs.

Traditional Visual Question Answering (VQA) techniques have made significant strides in the fields of image understanding and natural language processing [1,7,5,10]. By integrating deep learning models, they seamlessly merge images with natural language questions, thereby enabling an understanding of the questions and generation of corresponding answers. However, the application of traditional VQA models in the domain of medical imaging is subject to certain limitations. In radiological image diagnosis, current medical VQA methods primarily regard VQA as a classification problem over a finite answer set [4,8], which has struggled to effectively handle subtle differences between images and the open-ended nature of medical queries.

To bridge the gap between traditional VQA models and the diagnostic practices of radiologists, the concept of Difference Visual Question Answering (DiffVQA) has begun to be explored [6]. DiffVQA is better suited for radiologists' diagnostic practices primarily because it can effectively handle differences between images and simulate the longitudinal comparative analysis they conduct. Long-term comparative analysis is crucial in radiology for tracking disease progression, treatment response, and overall patient management.

Recently, EKAID [6] introduced the first medical DiffVQA dataset (namely MIMIC-Diff-VQA) and proposed a novel expert knowledge-aware graph representation learning model with the Assessment-Diagnosis-Intervention-Evaluation treatment procedure. PLURAL [3] introduces pretraining specifically tailored for the DiffVQA task using longitudinal chest X-ray data. However, both methods share a common limitation in treating paired image features, as they overlook the differences and relationships between images. They merely concatenate the features of the two images together. This simple concatenation approach renders these models less sensitive to the differences between images, whereas understanding the distinctions between images is precisely the key aspect of the DiffVQA task. Besides, the answers to DiffVQA questions are diverse and rich, making it difficult for traditional classification VQA to handle the overly large answer space. The generative model can solve this problem because it is not limited to a fixed set of answers. Recently, PMC-VQA [17] introduced a generative VQA model to medical VQA task for the first time and achieved a comparable performance to classification VQA methods.

In this paper, we propose a **Re**sidual **Al**ign framework (**ReAl**) for DiffVQA, designed to analyze chest X-ray images of the same patient captured at various time points and to offer insights into the discrepancies between the images, thus harmonizing with radiologists' diagnostic methodologies. Firstly, our ReAl model departs from the previous classification-based VQA paradigm and adopts a generative VQA paradigm to generate nuanced and flexible answers. Secondly, to heighten the model's emphasis on image disparities, we utilize the residual of two images as an additional input to ensure comprehensive understanding of the differences by the encoder. We also proposed a Residual Feature Align (RFA) module to ensure that our framework understands and learns the differences between corresponding images. Finally, we validate the performance of our model on the MIMIC-Diff-VQA dataset and compare it with previous state-of-the-art (SOTA) methods. Moreover, ablation experiments demonstrate that leveraging contrastive learning effectively focuses the model's attention on differences.

2 Methods

This section aims to introduce our proposed ReAl model, detailing its design and implementation. Firstly, we provide a definition of the DiffVQA problem and introduce the concept of generative VQA (Section 2.1). Subsequently, we explore the primary architecture of the ReAl model, encompassing its components and workflow (Section 2.2). Finally, we examine the RFA module to augment its emphasis on image disparities (Section 2.3).

2.1 Task Definition

In this subsection, we formally define the DiffVQA problem. We consider two images, labeled I_{ref} and I_{main}, typically representing chest X-ray images of the same patient taken at different time points. The objective is to address a series of

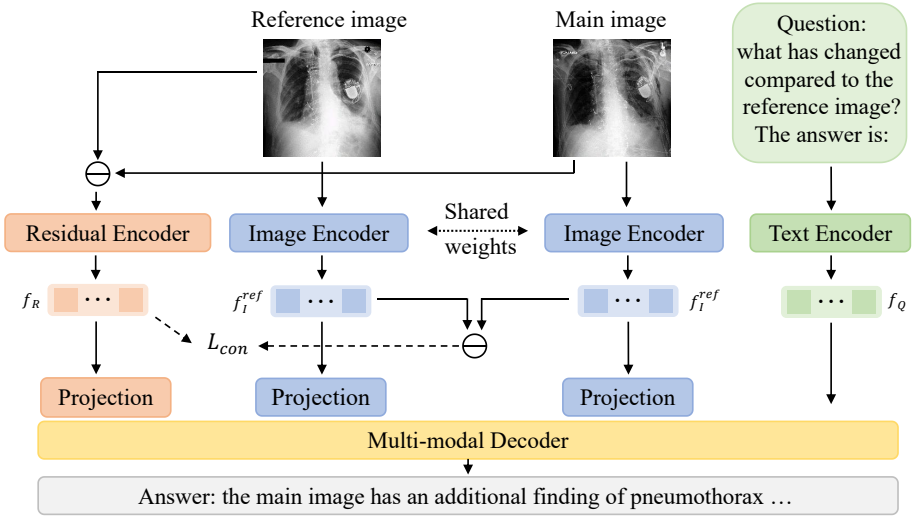

Fig. 2: Illustration of our proposed ReAl framework. The input of ReAl contains a main image, a reference image and associated question. ReAl additionally develops a residual encoder to acquire information of the difference between the main and reference images.

inquiries regarding changes in the patient's condition using these images. Let Q represent the question space, and A denote the answer space. For each question $q \in Q$, the aim is to devise a model that maps the question q and the image pair (I_{ref}, I_{main}) to an answer $a \in A$ within the answer space A.

In the current medical domain, VQA methods commonly adhere to a classification paradigm, where the answer space A is consistently a finite closed set, prompting their models to seek the optimal answer within this confined range. Nonetheless, the classification paradigm is ill-suited for DiffVQA, given that its answer space A typically remains open. Conversely, employing a generative model proves more suitable for the DiffVQA task, as such models possess the capability to dynamically generate diverse answers, thus facilitating a more nuanced analysis of subtle differences in medical images.

For a generative VQA model, the probability of generating a word w_t at each time step t can be represented as $p(w_t|w_{<t})$, where $w_{<t}$ denotes the sequence of words generated before the time step t. Assuming the target answer is $A = (a_1, a_2, ..., a_T)$, the loss function for the generative VQA task can be defined as the negative log-likelihood loss:

$$\mathcal{L}_{gen} = -\frac{1}{T}\sum_{t=1}^{T} \log p(a_t|a_{<t}),$$

where T is the length of the target answer. This loss function measures the discrepancy between the sequences generated by the model and the target answer,

aiming to minimize the loss and make the generated answer as close to the target answer as possible.

2.2 Architecture

In this section, we introduce the structure of the ReAl model. Specifically, the ReAl model primarily comprises multiple uni-modal encoders for feature extraction and a multimodal decoder for answer generation.

Uni-modal encoders In the ReAl model, we employ multiple uni-modal encoders, each serving a distinct purpose in processing different types of input data. Specifically, we utilize the following encoders: (1) the image encoder E_I, responsible for encoding visual information from reference and main images; (2) the text encoder E_T, designed to encode textual information from the question; and (3) the residual encoder E_R, dedicated to encoding temporal information to capture the sequential nature of longitudinal X-ray images.

To address the disparity between image and text embeddings, we apply a projection to the encoded features f_R, f_I^{ref} and f_I^{main} derived from the residual encoder E_R and the image encoder E_I, respectively. This additional measure guarantees the alignment of modalities in a congruent feature space, thereby facilitating seamless fusion and interaction in subsequent processing stages.

Multi-modal Decoder The main goal of the decoder is to generate answers using concatenated features extracted from previous stages. In this research, we chose GPT-2 as the decoder model owing to its robust performance in natural language generation tasks. Concretely, the decoder takes concatenated features from previous stages as input, comprising residual embeddings, image embeddings, and question embeddings. These concatenated features are then fed into the GPT-2 model for answer generation.

2.3 Residual Feature Alignment

In our methodology, we introduce the Residual Encoder, a novel component aimed at addressing the DiffVQA task, which involves comparing chest X-ray images of two patients to discern differences between them. Traditionally, only two encoders are utilized: one for image encoding and another for question encoding, extracting features from images and question, respectively. These features are subsequently concatenated and passed to a decoder for answer generation. However, our method deviates from this norm by incorporating an additional encoder, the residual encoder.

Residual Encoder shares a similar architecture to the preceding image encoder but with independent parameters. Our innovation lies in directly subtracting the reference image from the main image to obtain a residual image, which is

then fed into the residual encoder. Following this, we compute the disparity between features extracted from the two images and enforce this difference to align with the residual feature f_R outputted by the residual encoder using a loss function. This approach focuses the model's attention on image disparities, thereby enhancing DiffVQA task performance.

Specifically, the utilized consistency loss function can be denoted as:

$$L_{con} = ||f_I^{ref} - f_I^{main} - f_R||_2^2.$$

Here, f_I^{ref} and f_I^{main} denote the features of the reference and main images, respectively, while f_R represents the residual feature outputted by the residual encoder. The aim of this loss function is to minimize differences between image features, ensuring alignment of the difference between main and reference images with the residual feature outputted by the residual encoder.

2.4 Training and Inference

Training To train our ReAl, the generation loss $\mathcal{L}_{gen}$ and the consistent loss $\mathcal{L}_{con}$ are jointly leveraged, which can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \mathcal{L}_{con},$$

where $\mathcal{L}_{total}$ is the final loss used for optimization.

Inference The test procedure follows the same pipeline as the training, where a main image, a reference image and associated question are required. Based on the projected features associated to the text, images and residual image, the multi-modal decoder is responsible for producing the answer.

3 Experiments

3.1 Datasets

In our experimental section, we validate our model using the sole available DiffVQA dataset, MIMIC-Diff-VQA [6], which is derived from MIMIC-CXR. This dataset contains a total of 700,703 question-answer pairs extracted from 164,324 cases. It includes seven types of questions: abnormality, presence, view, location, type, level, and difference. However, for this study, we concentrate exclusively on the "difference" subset, comprising 131,563 questions related to differences. The dataset is partitioned into training, validation, and testing sets following an 8:1:1 ratio, which aligns with the official partitioning.

Table 1: Comparison of performance between previous SOTA methods and ReAl in the MIMIC-Diff-VQA dataset. The best results are shown in **bold**.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MCCFormers [13]	0.214	0.190	0.170	0.153	0.319	0.340	0.000
IDCPCL [16]	0.614	0.541	0.474	0.414	0.303	0.582	0.703
EKAID [6]	0.628	0.553	0.491	0.434	0.339	0.577	1.027
PLURAL [3]	0.704	0.633	0.575	0.520	0.381	0.653	1.832
ReAl (Ours)	**0.710**	**0.636**	**0.580**	**0.530**	**0.395**	**0.736**	**2.409**

Table 2: Results of the ablation experiments to investigate the impact of residual input and RFA module. The best results are in **bold**.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline	0.653	0.568	0.503	0.446	0.363	0.688	2.156
+ Residual input	0.698	0.613	0.547	0.489	0.388	0.732	2.346
+ RFA	**0.710**	**0.636**	**0.580**	**0.530**	**0.395**	**0.736**	**2.409**

3.2 Implementation and Evaluation

In the data preparation stage, we randomly cropped patches of size 512×512 from the X-rays as input. To mitigate overfitting of limited training data, we adjust brightness, contrast, saturation and sharpness to diversify the training data. During training, we choose ResNet-50 as the framework for both image and residual encoders to encode image features. The subsequent projection module consists of a transformer decoder and fully connected layers. The residual projection module have the same structure but separate parameters, while this two image projection modules are exactly the same. All codes are implemented in PyTorch-1.13 [12] and the models are trained on eight NVIDIA GeForce 3080Ti GPUs with 12GB memory.

We assess the framework's performance using prevalent natural language generation metrics, including BLEU [11], METEOR [2], ROUGE-L [9], and CIDEr-D [15]. These metrics serve as quantitative measures to evaluate the quality of the generated answers in terms of their similarity to reference answers, grammatical correctness, semantic relevance, and overall informativeness.

3.3 Comparison with State-of-the-arts

We conducted comparisons between our model and two prior state-of-the-art models, EKAID and PLURAL, employed in the DiffVQA task. In addition, we also compared our method with two state-of-the-art methods on image difference caption task, MCCFormers and IDCPCL. Image difference caption is a task similar to DiffVQA and has the same training goal, which is to find the differences between images and represent them with text output.

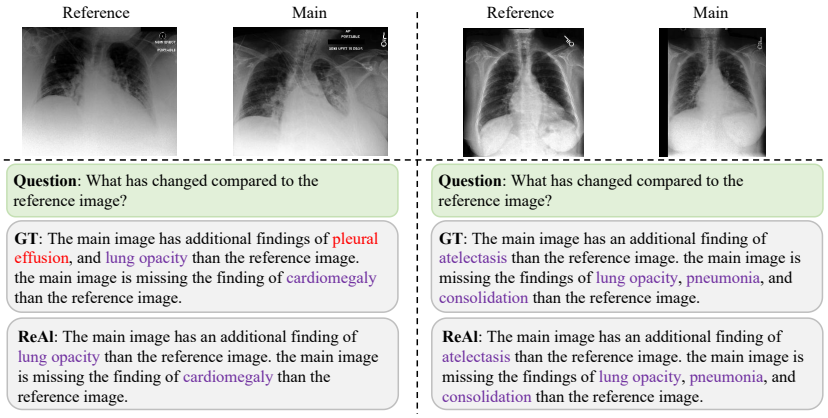

Fig. 3: Examples of difference questions and their corresponding answers generated by ReAl showcase its robust capability in identifying disparate regions and associating them with medical terminology. Red indicates incorrect or missing predictions, while purple represents correct predictions.

The table 1 presents the performance comparison on the MIMIC-Diff-VQA dataset. As indicated by the results, we can find: (1) our method consistently outperforms competitors on all metrics, with particularly significant enhancements observed in the ROUGE-L and CIDEr metrics. (2) Compared with PLURAL, which uses a large amount of paired X-Ray and report data to pre-training, our ReAl is more conducive to the model to focus on the changing regions without the need for pre-training. This is evidenced by higher performance improvements, with particularly significant enhancements observed in the ROUGE-L and CIDEr metrics. This result also demonstrates that ReAl pays more attention to the differences in the images, enabling it to generate more detailed answers to difference-related questions.

3.4 Ablation Studies

In this section, we conduct an ablation study to analyze the impact of different configurations of our model on its performance in the DiffVQA task. Specifically, we investigate three scenarios: the baseline model without residual input, the model with the addition of residual input (introducing the residual encoder branch), and the model with residual feature alignment based on the residual input. By comparing the performance of these three configurations, we aim to elucidate the individual contributions of residual input and residual feature alignment to the model's effectiveness in capturing image differences and generating accurate answers in the DiffVQA task. The outcomes of the ablation study suggest that incorporating residual input enhances the model's performance in the DiffVQA task. Moreover, the additional alignment of residual features further boosts performance. These findings underscore the positive influence of both

residual input and residual feature alignment on the model's capacity to capture image disparities and produce precise answers.

4 Conclusion

This paper proposes a novel method ReAl for medical difference visual question answering task. ReAl is not limited to the classification paradigm like existing medical VQA methods. The generative paradigm ReAl adopted can effectively generate more fine-grained answers. Thanks to the residual input and feature alignment, ReAl shows great potential for discovering differential change information. The experiments have demonstrated that the ReAl method outperformed previous SOTA methods in DiffVQA.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377, in part by Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, and in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Cho, Y., Kim, T., Shin, H., Cho, S., Shin, D.: Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. arXiv preprint arXiv:2402.08966 (2024)
4. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 64–74. Springer (2021)
5. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163 (2023)
6. 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. pp. 4156–4165 (2023)

7. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
8. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medial visual question answering. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
10. 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 p. 102611 (2023)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
13. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1971–1980 (2021)
14. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
15. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
16. Yao, L., Wang, W., Jin, Q.: Image difference captioning with pre-training and contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3108–3116 (2022)
17. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)