Region-Specific Retrieval Augmentation for Longitudinal Visual Question Answering: A Mix-and-Match Paradigm

Ka-Wai Yung[1] (✉), Jayaram Sivaraj[2], Danail Stoyanov[1], Stavros Loukogeorgakis[2], and Evangelos B. Mazomenos[1] (✉)

[1] UCL Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Department of Medical Physics and Biomedical Engineering, University College London, London, UK

[2] Department of Specialist Neonatal and Paediatric Surgery, Great Ormond Street Hospital NHS Foundation Trust, London, UK

{ka.yung.22,e.mazomenos}@ucl.ac.uk

Abstract. Visual Question Answering (VQA) has advanced in recent years, inspiring adaptations to radiology for medical diagnosis. Longitudinal VQA, which requires an understanding of changes in images over time, can further support patient monitoring and treatment decision-making. This work introduces RegioMix, a retrieval augmented paradigm for longitudinal VQA, formulating a novel approach that generates retrieval objects through a mix-and-match technique, utilizing different regions from various retrieved images. Furthermore, this process generates a pseudo-difference description based on the retrieved pair, by leveraging available reports from each retrieved region. To align such statements to both the posed question and input image pair, we introduce a Dual Alignment module. Experiments on the MIMIC-Diff-VQA X-ray dataset demonstrate our method's superiority, outperforming the state-of-the-art by 77.7 in CIDEr score and 8.3% in BLEU-4, while relying solely on the training dataset for retrieval, showcasing the effectiveness of our approach. Code is available at https://github.com/KawaiYung/RegioMix.

Keywords: Visual Question Answering · Retrieval Augmentation · X-ray Imaging.

1 Introduction

Visual Question Answering (VQA) represents a complex, multimodal task that combines the complexity of image analysis with natural language processing. In its core, VQA involves the fusion of visual and linguistic elements to provide a concise answer, in the form of natural language output, which necessitates a deep understanding of both the visual and language context of the question. Advancements in general domain VQA [1,10,2,6,7] have sparked efforts to transfer these methodologies to the medical field, for supporting medical diagnosis [28,14,18,25,17,23,20,24].

The task of Medical VQA (Med-VQA) is fundamentally more challenging, as the interpretation of medical images requires specialized knowledge, due to the presence of subtle yet clinically significant details. Many Med-VQA methods, however, consider only a single image at a time as the input. This approach overlooks the crucial aspect of the temporal dynamics of disease progression, which are pivotal for diagnosis, identifying risks of deterioration, treatment decision and outcome prediction [13,29]. To realize the full potential of Med-VQA it is important to develop models for longitudinal analysis, that can identify and textualize clinically relevant differences between medical images.

Longitudinal-VQA presents additional challenges to visual and textual understanding, as the model must also track and interpret changes across images over time. Hu *et al.* proposed EKAID [8], an image difference model for Longitudinal-VQA. Nevertheless, complex questions related to the identification of pathological differences between longitudinal images still pose challenges to the model, causing it to generate responses with incorrect findings.

We propose RegioMix, a novel approach for Longitudinal-VQA in Chest X-rays (CXRs), based on Retrieval Augmented Generation (RAG) [15]. RAG combines the strengths of retrieval-based and generative models, and dynamically retrieves relevant documents or data points from a vast knowledge base during the generation process. Due to its effectiveness, RAG has gained considerable attention across a variety of domains [3,9,27,16]. The application of RAG in this context presents unique challenges, as it requires the availability of a similar pair within the database that not only matches in pathology but also mirrors the specific changes or progression observed in the patient. Even when relaxing the constraint to include pairs from different patients, finding a closely matched retrieval between CXRs with appropriate pathology and severity remains difficult. To overcome these obstacles, we propose a Region-Specific RAG (RS-RAG) method through mix-and-matching. Exploiting the consistent nature of anatomical structures in CXRs, we generate retrieval objects by mixing-and-matching, based on their similarity, different regions from various images into an image pair. This approach enables the retrieved pair to closely match the query image pair while negating the need for an external CXR database.

RegioMix makes key contributions: **(1)** We introduce RS-RAG, a novel retrieval generation method achieved via mixing-and-matching combinations of regions from the training dataset. From the retrieved pair, we generate a pseudo-difference description that represents the changes between such pairs. **(2)** This description is often noisy and contains irrelevant findings. We design a Dual Alignment (DA) module to align this description with the input pair and the posed question. **(3)** We also propose PairNCE - a paired noise contrastive estimation loss to enhance the contextual understanding of difference representations. **(4)** In the MIMIC-Diff-VQA benchmark, RegioMix significantly outperforms the state-of-the-art by 77.7 in CIDEr score and 8.3% in BLEU-4, without relying on additional external databases or pretraining on natural images.

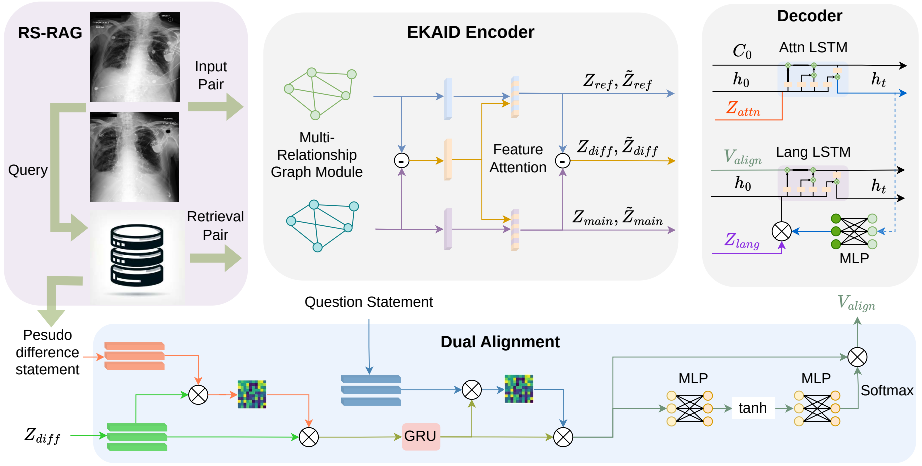


Fig. 1: The RegioMix architecture and workflow. The RS-RAG module generates a retrieval object closely matching the input pair, through mix-and-matching of different regions from various CXRs. RS-RAG also generates the pseudo-difference description based on the retrieved pair. This description is further aligned to the posed question and input pair difference through the Dual Alignment module. The aligned description, along with encoded features of the input and retrieval pair, are passed into the decoder for answer generation, where $Z_{lang} = [Z_{ref}; Z_{diff}; Z_{main}]$, $Z_{attn} = [Z_{ref}; Z_{diff}; Z_{main}; \tilde{Z}_{ref}; \tilde{Z}_{diff}; \tilde{Z}_{main}]$.

2 Method

Fig. 1 illustrates an overview of RegioMix, consisting of four key components: **(1)** A RS-RAG module for retrieving relevant region features from CXRs and constructing a pseudo-difference description of the changes between the retrieved image pair. **(2)** A DA module that aligns the pseudo-difference description with the input pair difference and the posed question. **(3)** An encoder backbone adapted from EKAID. **(4)** A decoder built from two Long Short Term Memory (LSTM) modules for generating the corresponding answer.

Region-Specific Retrieval Augmentation: We consider a pair of input X-ray images from the same patient, taken at different times. X_{ref} denotes the reference preceding image, while X_{main} is the more recent image. Following the EKAID methodology, the anatomical region and the disease region within the two images are extracted using Faster-RCNN [22]. Performing retrieval at paired image level is often complicated by the scarcity of similar paired X-rays in datasets. Allowing pairings from different patients does alleviate some constraints, yet it remains suboptimal due to the infrequency of images sharing identical pathologies and severity. Contrary to image difference captioning in general domain VQA, where the overall scene provides crucial relational information about objects that have undergone change, longitudinal CXR analysis requires a

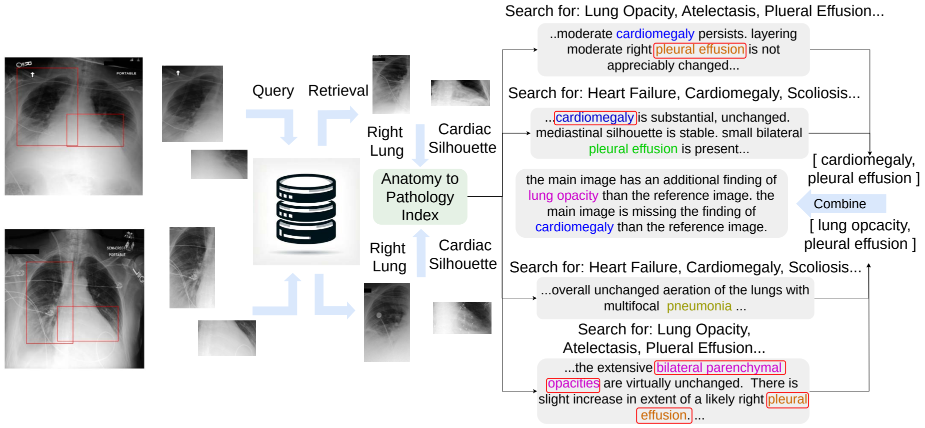

Query
Retrieval
Right Lung
Cardiac Silhouette
Anatomy to Pathology Index
Right Lung
Cardiac Silhouette
Search for: Lung Opacity, Atelectasis, Plueral Effusion...
..moderate cardiomegaly persists. layering moderate right pleural effusion is not appreciably changed...
Search for: Heart Failure, Cardiomegaly, Scoliosis...
...cardiomegaly is substantial, unchanged. mediastinal silhouette is stable. small bilateral pleural effusion is present...
the main image has an additional finding of lung opacity than the reference image. the main image is missing the finding of cardiomegaly than the reference image.
[cardiomegaly, pleural effusion]
Combine
[lung opcacity, pleural effusion]
Search for: Heart Failure, Cardiomegaly, Scoliosis...
...overall unchanged aeration of the lungs with multifocal pneumonia ...
Search for: Lung Opacity, Atelectasis, Plueral Effusion...
...the extensive bilateral parenchymal opacities are virtually unchanged. There is slight increase in extent of a likely right pleural effusion. ...

Fig. 2: Generation pipeline of the pseudo-difference description. For each retrieved region, we identify all pathologies listed in the corresponding report that could be present in that anatomical region, creating a list of findings for each query image. The two lists are then synthesized into a coherent sentence, forming the pseudo-difference description.[1]

different approach. Owing to the consistent nature of anatomical structures in X-rays, the general scene contributes less to discerning changes. The focus instead shifts to the specific alterations within these structures, as the overall anatomy is typically common, thus uniform and predictable. We therefore propose constructing pseudo-retrieval pairs at the region level through mix-and-matching. This approach allows each retrieved region to be sourced from different images and patients, effectively mitigating the issue of limited paired examples in the dataset, while ensuring the retrieved pair representations are closely matched to the input pair.

Given anatomical region features set $Z_a \in \mathbb{R}^{M,D}$ and disease region features set $Z_d \in \mathbb{R}^{M,D}$, with M the number of top detected regions and D the feature dimension, we perform feature-wise concatenation resulting in a combined region feature set $Z_c = [Z_a; Z_d] \in \mathbb{R}^{M,2D}$. The retrieval process involves comparing each individual combined region feature $z_c \in \mathbb{R}^{2D}$ with all other z_c from the training dataset. We perform this with cosine similarity and select the top K matches for each comparison. Consequently, for each query image, we acquire a set of retrieved region features denoted as $Z_c^R \in \mathbb{R}^{M,K,2D}$. The input and retrieval pairs are processed through an encoder, derived from the EKAID framework, which integrates a Graphical Module for computing the spatial, semantic, and implicit relationships within the image, alongside a Feature Attention module for feature fusion. The encoder outputs feature-sets of the reference image $Z_{ref}, \tilde{Z}_{ref} \in \mathbb{R}^D$, main image $Z_{main}, \tilde{Z}_{main} \in \mathbb{R}^D$, and the difference feature $Z_{diff}, \tilde{Z}_{diff} \in \mathbb{R}^D$ are computed. Here, $\tilde{Z}$ represents features from the retrieval pair.

[1] See supplementary Table S1 for the full Anatomy-to-Pathology index search list.

Fig. 2 illustrates the procedure for retrieval and generation of pseudo-difference descriptions from the retrieved features. For each retrieved region feature $z_c^R \in \mathbb{R}^{2D}$, we first derive its predicted anatomical label as identified by the Faster R-CNN, along with the full diagnostic report associated with the corresponding image. From the anatomic label, we extract all relevant findings noted within the report that belong to the identified anatomical region through a search list. Consequently, we generate a list of findings for each corresponding image of the retrieved pair. Finally, the two lists are transformed into a coherent sentence that articulates the pseudo-differences description.

Dual Alignment (DA) Module: The process of generating pseudo-difference descriptions can yield noisy results with extraneous findings that are not directly relevant to the actual changes of the input pair. Additionally, not all questions are directly related to the changes between the reference and main images. As a result, the information derived from the pseudo-difference descriptions may not always be beneficial or directly applicable.

To address this, we introduce a DA module (see Fig. 1 bottom). Given the embedded pseudo-difference description $V_p \in \mathbb{R}^{N,D}$, where N is the sequence length and D is the embedding dimension, we first align V_p with the input difference Z_{diff} as formulated in Eq. 1. The aligned output V_{da} is then aligned with the question V_q as indicated in Eq. 2. Finally, self-attention is performed on the aligned embedding V_{qda} to compute module output V_{align} (Eq. (3)):

$$V_{da} = \text{GRU}(\text{softmax}\left(\frac{(V_p W^Q)(Z_{diff} W^K)^T}{\sqrt{d_k}}\right)(Z_{diff} W^V)) \tag{1}$$

$$V_{qda} = \text{softmax}\left(\frac{(V_q W^Q)(V_{da} W^K)^T}{\sqrt{d_k}}\right)(V_{da} W^V) \tag{2}$$

$$V_{align} = \text{MLP}(\tanh(\text{MLP}(V_{qda})))(V_{qda}) \tag{3}$$

W^Q, W^K and W^V are weight matrices, d_k is the scaling factor, GRU and MLP denote the Gated Recurrent Unit [4] and Multi-Layer Perceptron respectively.

Answer Generation: To decode features from encoder into a corresponding answer, we employ an Attention LSTM module and a Language LSTM module, shown on the right of Fig. 1. In the Attention LSTM, all input pair and retrieval pair features are concatenated together and used as input to compute the attention weights $\alpha \in \mathbb{R}^3$. Subsequently, the calculated α are multiplied and summed with the concatenation of the three non-retrieval features $[Z_{ref}; Z_{main}; Z_{diff}] \in \mathbb{R}^{3,D}$, and passed into the Language LSTM. We initialize the cell state C_0 of the Language LSTM with the aligned-pesudo difference description V_{align}, and generate answer with the weighted sum of features.

Pair-wise Contrastive Learning: To strengthen the difference representations learnt from the encoder, we treat the input pair difference embedding Z_{diff} and corresponding retrieval pair difference embedding $\tilde{Z}_{diff}$ as a positive pair, and employ the InfoNCE loss (Eq. 4), where τ is the temperature parameter [19], to

Table 1: Performance comparison of existing Diff-VQA methods to proposed RegioMix based on *difference* questions and *all* questions.

	Difference				All	
Metrics	MCCFormers [21]	IDCPCL [26]	EKAID	RegioMix	EKAID	RegioMix
BLEU-1	21.4	61.4	62.8	**70.5**	62.4	**70.4**
BLEU-2	19.0	54.1	55.3	**63.3**	54.1	**62.3**
BLEU-3	17.0	47.4	49.1	**57.2**	47.7	**56.0**
BLEU-4	15.3	41.4	43.4	**51.7**	42.2	**50.4**
METEOR	31.9	30.3	33.9	**38.1**	33.7	**38.3**
ROUGE-L	34.0	58.2	57.7	**65.1**	64.5	**72.6**
CIDEr	0	70.3	102.7	**180.4**	189.3	**235.4**

maximize the agreement between them. The total loss function (Eq. 5) is:

$$\mathcal{L}_{\text{PairNCE}} = -\log \frac{\exp(\text{sim}(Z^i_{diff}), (\tilde{Z}^i_{diff}))/\tau)}{\sum_{j=1}^{n} \exp(\text{sim}(Z^i_{diff}, \tilde{Z}^j_{diff}))/\tau)} \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{\text{speaker}} + \lambda \mathcal{L}_{\text{PairNCE}} \tag{5}$$

where $\mathcal{L}_{\text{speaker}}$ refers to the negative log-likelihood loss between the generated answer and ground truth, and λ is a trade-off parameter that adjusts the balance between the speaker loss and the PairNCE loss.

3 Results and Discussion

Dataset and Settings: We use the publicly available MIMIC-Diff-VQA[2] dataset which is built from the MIMIC-CXR [8,11] dataset and contains 700,703 questions with JPG images. There are seven types of questions: 1) Difference (23%) 2) Presence (22%) 3) Abnormality (21%) 4) Location (12%) 5) Level (10%) 6) View (8%) 7) Type (4%). For retrieval, we use FAISS [12] for efficient searching. We empirically set, number of top retrievals $K = 5$ and trade-off parameter $\lambda = 0.5$ for PairNCE. When applying Faster-RCNN, we follow EKAID and downsample to 1K keeping aspect ratio for efficient computation. We keep the rest of the hyper-parameters including the learning rate and seeding to be the same as EKAID. We do not perform pre-training with natural images and train directly with MIMIC-Diff-VQA.
Results: Table 1 presents a performance comparison between RegioMix and other existing models. Notably, RegioMix demonstrates superior performance across all metrics when compared to the state-of-the-art EKAID for both *all* and *difference* question types.

Particularly, RegioMix shows a notable improvement of 77.7 points (+76%) of CIDEr score in *difference* questions compared to EKAID. The CIDEr metric measures the similarity of the generated answer to the ground truth. Unlike

[2] https://physionet.org/content/medical-diff-vqa/1.0.0/

Table 2: Results breakdown by question type.

	EKAID				RegioMix (Ours)			
Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Location	52.5	36.4	21.0	14.4	**61.1**	**47.0**	**34.5**	**26.9**
Abnormality	48.2	33.3	19.7	10.9	**58.2**	**43.7**	**30.4**	**19.9**
Presence	80.1	-	-	-	**85.0**	-	-	-
View	94.8	94.1	-	-	**98.7**	**98.4**	-	-
Level	49.6	-	-	-	**60.1**	-	-	-

Table 3: (Left) Accuracy comparison on *non-difference* questions. (Right) Ablation results of the RegioMix architecture components.[3]

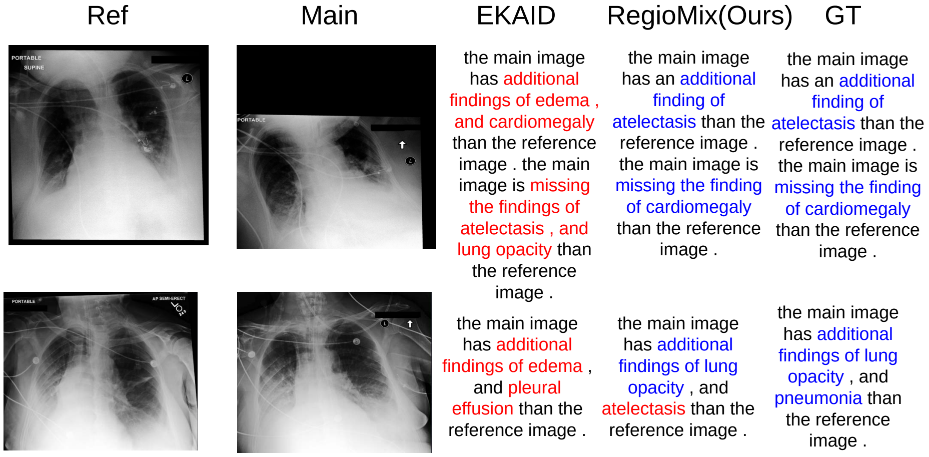
Question	Open	Closed	Total
MMQ [5]	11.5	10.8	11.5
EKAID	26.4	79.9	52.5
Ours	**35.9**	**85.7**	**60.2**

Metrics	EKAID	+RS-RAG	+DA	+$L_{pairNCE}$
BLEU-1	62.8	69.9	70.4	**70.5**
BLEU-2	55.3	62.4	63.1	**63.3**
BLEU-3	49.1	56.2	57.1	**57.2**
BLEU-4	43.4	50.4	51.5	**51.7**
METEOR	33.9	**38.1**	**38.1**	**38.1**
ROUGE-L	57.7	65.1	**65.3**	65.1
CIDEr	102.7	171.5	**180.4**	**180.4**

BLEU, which applies uniform n-gram weighting, CIDEr places greater emphasis on rare but informative n-grams by assigning a higher weight while reducing the influence of frequent but less informative ones. This aspect is crucial in the domain of Med-VQA, where answers often adhere to a general template but contain key, rare findings. For example, in the response "the main image has additional findings of *pleural effusion*". Achieving a satisfactory BLEU score might be possible with a semantically similar but clinically inaccurate answer such as "the main image has additional findings of *pleural thickening*" based solely on n-gram matches. However, such an answer is clinically not satisfactory as it fails to provide correct information about the images. The significant improvement in CIDEr indicates that RegioMix excels not only in n-gram level accuracy but also in generating the correct, clinically relevant rare words, which often correspond to crucial findings.

RegioMix also achieves substantial improvements in other metrics with a 8.3% increase in BLEU-4 and 7.4% in ROUGE-L for *difference* questions, and a 8.2% increase in BLEU-4 (95% CI [48.6-51.1]) and 8.1% increase in ROUGE-L (95% CI [234.0-237.0]) for *all* questions. The higher BLEU-4 scores suggest that our proposed method generates answers with better overall quality and structure, indicating that it is not merely overfitting to more frequent common words. This comprehensive improvement across multiple metrics underscores the effectiveness of our approach. Table 2 shows a performance breakdown by

[3] See supplementary Table S2 for ablation studies on K and λ.



EKAID	RegioMix(Ours)	GT
the main image has additional findings of edema , and cardiomegaly than the reference image . the main image is missing the findings of atelectasis , and lung opacity than the reference image .	the main image has an additional finding of atelectasis than the reference image . the main image is missing the finding of cardiomegaly than the reference image .	the main image has an additional finding of atelectasis than the reference image . the main image is missing the finding of cardiomegaly than the reference image .
the main image has additional findings of edema , and pleural effusion than the reference image .	the main image has additional findings of lung opacity , and atelectasis than the reference image .	the main image has additional findings of lung opacity , and pneumonia than the reference image .

Fig. 3: Examples of answers generated by RegioMix compared to EKAID from question 'what has changed compared to the reference image?'. Blue text indicates correct and red incorrect findings. RegioMix can determine the changes in the input pair and generates more accurate responses, while EKAID gives answers with more missing or incorrect findings.[4]

question type. RegioMix demonstrates superiority, compared to EKAID, over all types of questions, and specifically a BLEU-4 improvement of 9.0% and 12.5% in *abnormality* and *location* questions respectively. Compared to EKAID on *non-difference* questions (Table 3 left), RegioMix achieves improvements of 9.5% in *open* questions and 5.8% in *closed* questions.

For difference questions, RegioMix produced 1512 exact matches compared to 874 in EKAID, comparsion examples are shown in Fig. 3. RegioMix generates response with more accurate pathologies, while EKAID generates responses with incorrect or missing findings.

Ablation Studies: Table 3 (right) presents results from an ablation study on the RegioMix architecture components. The inclusion of the RS-RAG module produces a noticeable increase of 68.8 in CIDEr score and 7.0% in BLEU-4, indicating enhanced accuracy in generating clinically relevant rare words. However, the pseudo-difference description, while generally beneficial, can lead to misleading information when the posed question is irrelevant to the identified differences. With the DA module, the BLEU score further increases by 1.1% along with a 8.9 increase in the CIDEr score, underscoring the importance of correctly aligning the pseudo-difference description to the specific context of the question. Furthermore, the PairNCE loss contributes to additional improvements in the BLEU score without affecting the METEOR and CIDEr scores.

[4] See supplementary Figure S1 and Video S1 for additional results on image pairs.

4 Conclusion

This paper introduces RegioMix, a novel retrieval-augmented pipeline for Longitudinal VQA in CXRs. We generate retrieval objects through a mix-and-match approach with regions from various CXRs, along with a pseudo-difference description of the clinical changes between them. The introduction of the Dual Alignment module ensures the alignment of the description to the posed questions and input image pairs. Experimentation on the MIMIC-Diff-VQA benchmark, RegioMix significantly outperforms the current state-of-the-art, achieving a 77.7 increase in CIDEr score and a 8.3% improvement in BLEU-4 for difference-focused questions, alongside a 46.1 increase in CIDEr score and 8.2% improvement in BLEU-4 across all question types, respectively.

Acknowledgments. This work was funded in whole, or in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z, NS/A000050/1], the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies [CiET1819/2/36]. For the purpose of open access, the authors have applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

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

References

1. Agrawal, A., Lu, J., Antol, S., et al.: Vqa: Visual question answering. Int. J. Comput. Vis. **123**, 4 – 31 (2015)
2. Chen, L., Yan, X., Xiao, J., et al.: Counterfactual samples synthesizing for robust visual question answering. CVPR pp. 10797–10806 (2020)
3. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. In: ICLR (2023)
4. Cho, K., van Merrienboer, B., Çaglar Gülçehre, et al.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: EMNLP (2014)
5. Do, T., Nguyen, B.X., Tjiputra, E., et al.: Multiple meta-model quantifying for medical visual question answering. In: MICCAI (2021)
6. Gao, F., Ping, Q., Thattai, G., et al.: Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. CVPR pp. 5057–5067 (2022)
7. Gokhale, T., Banerjee, P., Baral, C., et al.: Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In: EMNLP (2020)
8. Hu, X., Gu, L., An, Q., et al.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In: KDD. p. 4156–4165 (2023)
9. Izacard, G., Lewis, P.S.H., Lomeli, M., et al.: Atlas: Few-shot learning with retrieval augmented language models. J. Mach. Learn. Res. **24**, 251:1–251:43 (2023)
10. Jiang, H., Misra, I., Rohrbach, M., et al.: In defense of grid features for visual question answering. CVPR pp. 10264–10273 (2020)

11. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. data **6** (2019)
12. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
13. Karwande, G., Mbakawe, A., Wu, J.T., et al.: Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In: MICCAI. vol. 13431, pp. 581–591 (2022)
14. Khare, Y., Bagal, V., Mathew, M., et al.: Mmbert: Multimodal bert pretraining for improved medical vqa. ISBI pp. 1033–1036 (2021)
15. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NIPS (2020)
16. Lin, B.Y., Tan, K., Miller, C., et al.: Unsupervised cross-task generalization via retrieval augmentation. NIPS **35**, 22003–22017 (2022)
17. Liu, B., Zhan, L.M., Wu, X.M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: MICCAI (2021)
18. Liu, B., Zhan, L.M., Xu, L., et al.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. ISBI pp. 1650–1654 (2021)
19. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018)
20. Pellegrini, C., Keicher, M., Özsoy, E., other: Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In: MICCAI. pp. 409–419 (2023)
21. Qiu, Y., Yamamoto, S., Nakashima, K., et al.: Describing and localizing multiple changes with transformers. In: ICCV. pp. 1951–1960 (2021)
22. Ren, S., He, K., Girshick, R.B., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2015)
23. van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., et al.: Open-ended medical visual question answering through prefix tuning of language models. In: MICCAI. pp. 726–736 (2023)
24. Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Localized questions in medical visual question answering. In: MICCAI. pp. 361–370 (2023)
25. Vu, M.H., Löfstedt, T., Nyholm, T., et al.: A question-centric model for visual question answering in medical imaging. IEEE Trans Med Imaging **39**, 2856–2868 (2020)
26. Yao, L., Wang, W., Jin, Q.: Image difference captioning with pre-training and contrastive learning. In: AAAI (2022)
27. Zakka, C., Shad, R., Chaurasia, A., et al.: Almanac—retrieval-augmented language models for clinical medicine. NEJM AI **1**(2), AIoa2300068 (2024)
28. Zhan, L.M., Liu, B., Fan, L., et al.: Medical visual question answering via conditional reasoning. ACM-MM (2020)
29. Zhu, Q., Mathai, T.S., Mukherjee, P., et al.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In: MICCAI. vol. 14224, pp. 189–198 (2023)