Leveraging Coarse-to-Fine Grained Representations in Contrastive Learning for Differential Medical Visual Question Answering

Xiao Liang[1], Yin Wang[1], Di Wang[1] [1(✉)], Zhicheng Jiao[2], Haodi Zhong[1], Mengyu Yang[3], and Quan Wang[1]

[1] Xidian University, Xi'an, China
wangdi@xidian.edu.cn
[2] Warren Alpert Medical School of Brown University, Providence RI 02903, USA
[3] Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Chest X-ray Differential Medical Visual Question Answering (Diff-MedVQA) is a novel multi-modal task designed to answer questions about diseases, especially their differences, based on a main image and a reference image. Compared to the widely explored visual question answering in the general domain, Diff-MedVQA presents two unique issues: (1) variations in medical images are often subtle, and (2) it is impossible for two chest X-rays taken at different times to be at exactly the same view. These issues significantly hinder the ability to answer questions about medical image differences accurately. To address this, we introduce a two-stage framework featuring Coarse-to-Fine Granularity Contrastive Learning. Specifically, our method initially employs an anatomical encoder and a disease classifier to obtain fine-grained visual features of main and reference images. It then integrates the anatomical knowledge graph to strengthen the relationship between anatomical and disease regions, while Multi-Change Captioning transformers identify the subtle differences between main and reference features. During pre-training, Coarse-to-Fine Granularity Contrastive Learning is used to align knowledge enhanced visual differences with keyword features like anatomical parts, symptoms, and diseases. During the Diff-MedVQA fine-tuning, the model treats the differential features as context-grounded queries, with language modeling guiding answer generation. Extensive experiments on the MIMIC-CXR-Diff dataset validate the effectiveness of our proposed method. Code is available at https://github.com/big-white-rabbit/Coarse-to-Fine-Grained-Contrastive-Learning.

Keywords: Differential Medical Visual Question Answering · Medical Visual Question Answering · Change Captioning · Visual and Language.

1 Introduction

The medical community actively develops deep learning algorithms for Chest X-ray data analysis, as seen in MIMIC-CXR [9], IU-Xray [11], and Chexpert [8].

Such work includes disease classification [3], report generation [7, 14, 21], and answering medically-related questions [2,12]. Among these, Differential Medical Visual Question Answering (Diff-MedVQA) [6] aligns closely with radiologist practices. It involves comparing current and prior Chest X-rays of the same patient to assess changes in anatomical structures and disease progression. This necessitates a model capable of accepting the main image and reference image as input, understanding the differences between them, and more importantly, verifying whether these differences align with expectations based on the doctor's question. Despite Diff-MedVQA's significant practical importance, this task has not yet been extensively investigated by researchers. Fortunately, in the general domain, Change Captioning [17, 22], which identifies differences between two images, has inspired our work. Typical methods like Multi-Change Captioning transformers(MCCFormer) [18] employ two identical image encoders to encode features of the main and reference images, utilize a transformer to capture their differences, and feed these differential features into a decoder to generate a caption. However, compared to general domain images, Diff-MedVQA faces two unique challenges: (1) Changes in medical images are often subtle, meaning that global image encoders such as ResNet [5] struggle to specifically capture the areas of difference. (2) In Change Captioning, the two images are typically captured in a single shot, ensuring an identical perspective. However, two Chest X-rays taken at different times cannot have perfectly matching viewpoints, adding complexity to the comparison of image differences.

To address these challenges, we introduce a two-stage framework with Coarse-to-Fine Granularity Contrastive Learning. Initially, we obtain fine-grained visual features of main and reference images using the anatomical encoder and the disease classifier. Visual region relationships are enhanced using the anatomical-symptom knowledge graph, followed by MCCFormer identifying the subtle visual differences between the main and reference features. During the pre-training stage, Coarse-to-Fine Granularity Contrastive Learning is used to align these visual differences with keyword textual features such as anatomical parts, symptoms, and diseases, combined with Image-Text Matching and Masked Language Modeling for text encoder training. In the Diff-MedVQA Fine-tuning stage, the model uses the visual differential features as context-grounded query inputs, with Language Modeling steering answer generation. Our contributions are summarized as follows:

- We propose a novel two-stage training framework for Diff-MedVQA with Coarse-to-Fine Granularity Contrastive Learning, demonstrating the feasibility of aligning image differential features with textual features.
- This method employs an anatomical knowledge graph to model the relationships between fine-grained anatomical features of Chest X-rays and corresponding diseases, achieving more robust differential representations.
- Our approach achieves state-of-the-art performance on the MIMIC-CXR-Diff [6] dataset, demonstrating the benefits of fine-grained contrastive learning and anatomical knowledge graph enhancement.

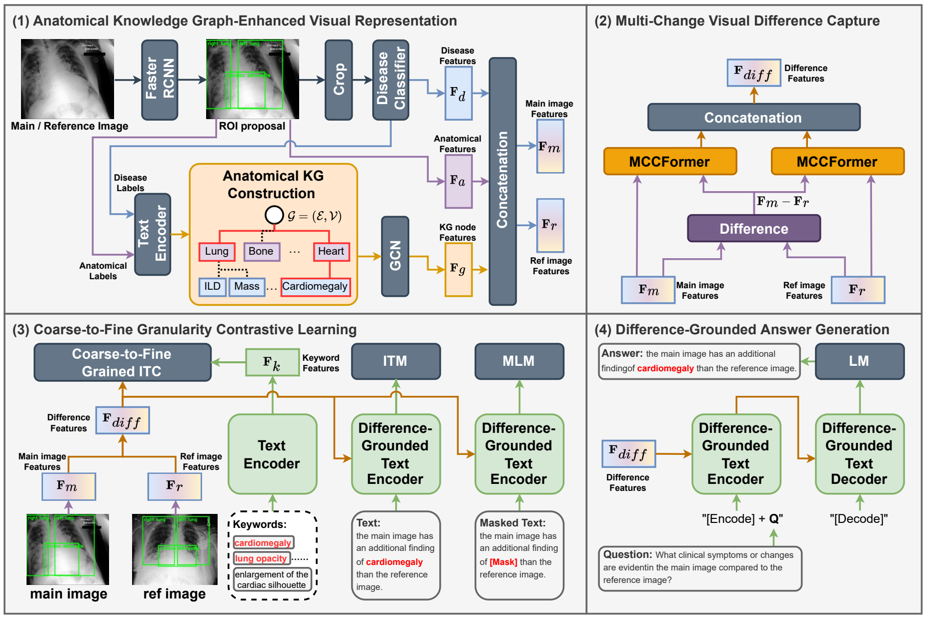

(1) Anatomical Knowledge Graph-Enhanced Visual Representation
Main / Reference Image
Faster RCNN
ROI proposal
Crop
Disease Classifier
Disease Features
$\mathbf{F}_d$
Main image Features
Anatomical Features
$\mathbf{F}_m$
$\mathbf{F}_a$
Concatenation
Disease Labels
Text Encoder
Anatomical Labels
Anatomical KG Construction
$\mathcal{G} = (\mathcal{E}, \mathcal{V})$
Lung
Bone
Heart
ILD
Mass
Cardiomegaly
GCN
KG node Features
$\mathbf{F}_g$
$\mathbf{F}_r$
Ref image Features
(2) Multi-Change Visual Difference Capture
$\mathbf{F}_{diff}$
Difference Features
Concatenation
MCCFormer
MCCFormer
$\mathbf{F}_m - \mathbf{F}_r$
Difference
$\mathbf{F}_m$
Main image Features
Ref image Features
$\mathbf{F}_r$
(3) Coarse-to-Fine Granularity Contrastive Learning
Coarse-to-Fine Grained ITC
$\mathbf{F}_k$
Keyword Features
ITM
MLM
Difference Features
$\mathbf{F}_{diff}$
Main image Features
Ref image Features
$\mathbf{F}_m$
$\mathbf{F}_r$
Text Encoder
Difference-Grounded Text Encoder
Difference-Grounded Text Encoder
main image
ref image
Keywords:
cardiomegaly
lung opacity
enlargement of the cardiac silhouette
Text: the main image has an additional finding of cardiomegaly than the reference image.
Masked Text: the main image has an additional finding of [Mask] than the reference image.
(4) Difference-Grounded Answer Generation
Answer: the main image has an additional findingof cardiomegaly than the reference image.
LM
$\mathbf{F}_{diff}$
Difference Features
Difference-Grounded Text Encoder
Difference-Grounded Text Decoder
"[Encode] + Q"
"[Decode]"
Question: What clinical symptoms or changes are evidentin the main image compared to the reference image?

Fig. 1. Overview of our framework, including Anatomical Knowledge Graph-Enhanced Visual Representation, Multi-Change Visual Difference Capture, Coarse-to-Fine Granularity Contrastive Learning, and Difference-Grounded Answer Generation.

2 Methodology

2.1 Problem Definition

Given a pair of Chest X-ray images $\mathcal{I}_m, \mathcal{I}_r$ taken at different times of the same patient, Diff-MedVQA aims to generate an answer a for a given question q:

$$\hat{a} = \arg\min \mathcal{F}_\theta(a \,|\, \mathcal{I}_m, \mathcal{I}_r, q) \tag{1}$$

Our framework comprises four steps: Anatomical Knowledge Graph-Enhanced Visual Representation, Multi-Change Visual Difference Capture, Coarse-to-Fine Granularity Contrastive Learning, and Difference-Grounded Answer Generation, detailed in Figure 1 and subsequent sections.

2.2 Anatomical Knowledge Graph Enhanced Visual Representation

Diff-MedVQA has to handle Chest X-ray pairs that often vary in their imaging angles. This variation renders differential features sensitive to pixel-level shifts when employing identical image encoders for both images, thus complicating the precise detection of anatomical symptoms or diseases. To obtain a more robust visual representation against varying perspectives, we introduced a Chest X-ray anatomical knowledge graph, which contains each anatomical part and its potential corresponding symptoms. To construct this graph, we first used FasterRCNN [20], pre-trained on the Chest ImaGenome [23], to detect anatomical Regions of Interest (ROIs) and encode their features $\mathbf{F}_a = \{f_{a_i} \in \mathbb{R}^d\}_{i=1}^n$, with n representing the number of ROIs. Then, a ResNet pre-trained on VinDr-CXR [15] is used to extract disease features $\mathbf{F}_d = \{f_{d_i} \in \mathbb{R}^d\}_{i=1}^n$ from these ROIs for multi-label disease classification. Based on the anatomical location and disease abnormality relationships provided by Chest ImaGenome, we construct the Anatomical Knowledge Graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$. All anatomical location labels and their corresponding disease labels are encoded into node features $v_i \in \mathbb{R}^d$ by the text encoder, resulting in $\mathcal{V} = \{v_i\}_{i=1}^N$. Here, N represents the total number of anatomical locations and diseases as detailed in [25]. If a disease is present in a certain ROI, their relationship is represented by an edge e_j, where $\mathcal{E} = \{e_j\}_{j=1}^M$, and M denotes the number of edges. A Graph Convolutional Network (GCN) [10] is employed to model the Anatomical Knowledge Graph. The graph convolution on can be expressed as:

$$\mathbf{H}^{(l)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)}) \tag{2}$$

Here, $\mathbf{H}^{(l)}$ is the node feature matrix at layer l, with $\mathbf{H}^{(0)}$ originating from node embeddings $\mathcal{V} \in \mathbb{R}^{N \times d}$. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, from edge set $\mathcal{E}$ of $\mathcal{G}$, normalizes to $\hat{\mathbf{A}}$. $\mathbf{W}^{(l)} \in \mathbb{R}^{d \times d}$ is the learnable weight matrix for each layer, and σ represents a non-linear activation function like ReLU. After 2 GCN layers, node features $\mathbf{H}^{(l)}$ are refined into angle-robust anatomical representations $\mathbf{F}_g \in \mathbb{R}^{N \times d}$. These are then concatenated with ROI and disease features $\mathbf{F}_a$ and $\mathbf{F}_d$ to form the main and reference image representations, $\mathbf{F}_m$ and $\mathbf{F}_r$, which reside in $\mathbb{R}^{(2n+N) \times d}$ for difference capture.

2.3 Multi-Change Visual Difference Capture

To compare differences between main and reference images, we use a symmetrical Multi-Change Capture Transformer (MCCFormer) to process the main and reference visual representations $\mathbf{F}_m$ and $\mathbf{F}_r$:

$$\mathbf{F}_{diff} = [\text{MCCFormer}(\mathbf{F}_m, \mathbf{F}_m - \mathbf{F}_r); \text{MCCFormer}(\mathbf{F}_r, \mathbf{F}_m - \mathbf{F}_r)] \tag{3}$$

where [;] denotes concatenation. The MCCFormer can be expressed as:

$$\text{MCCFormer}(\mathbf{F}_1, \mathbf{F}_2) = \text{MHSA}(\mathbf{F}_1 + \mathbf{E}_{pos1}, \mathbf{F}_2 + \mathbf{E}_{pos2}) \tag{4}$$

Here, Multi-Head Self-Attention (MHSA) utilizes positional embeddings $\mathbf{E}_{pos1}$ and $\mathbf{E}_{pos2}$ for the main and reference image visual representations to precisely preserve the differences' positional information.

2.4 Coarse-to-Fine Grained Contrastive Learning

Although difference features $\mathbf{F}_{diff} \in \mathbb{R}^{V \times d}$, where $V = 4n + 2N$, capture the visual discrepancies between image pairs, precise text generation necessitates modal alignment between vision and text. Therefore, we apply Coarse-to-Fine Grained Contrastive Learning in the pre-training stage to enhance semantic alignment between $\mathbf{F}_{diff}$ and text features of keywords in answers, like diseases and symptoms. Specifically, PubMedBERT [4] serves as the text encoder to extract the feature $\mathbf{F}_k$ for each keyword, with W representing the number of text tokens. Then, we average $\mathbf{F}_{diff}$ and $\mathbf{F}_k$ to form the global features f_{diff} and f_k, and subsequently calculate their cosine similarity:

$$s_g = \frac{\langle f_{diff}, f_k \rangle}{\|f_{diff}\| \|f_k\|} \tag{5}$$

Acknowledging that image pairs may correspond to multiple keywords, we use a sigmoid function instead of the traditional softmax for logits computation in image-text contrastive learning. Binary cross-entropy (BCE) is employed for optimization. Consequently, the global contrastive loss is:

$$\mathcal{L}_{\text{global}} = -\sum_{i=1}^{B} \left(y_i \log(\sigma(s_{g_i})/\tau_1) + (1 - y_i) \log(1 - \sigma(s_{g_i})/\tau_1) \right), \tag{6}$$

where B is the batch size, y_i indicates the relevance of the keyword to the i-th image pair, σ denotes the sigmoid function, and τ_1 is the global temperature parameter. Given that difference feature $\mathbf{F}_{diff}$ contains varied anatomical parts and disease information, with each part corresponding to different keywords, global contrastive learning is insufficient to capture these nuanced relationships. Therefore, we calculate fine-grained similarity between $\mathbf{F}_{diff}$ and $\mathbf{F}_k$:

$$\mathbf{s} = \mathbf{F}_k \mathbf{F}_{diff}^T \tag{7}$$

where the similarity matrix $\mathbf{s} \in \mathbb{R}^{W \times V}$ is employed to calculate the attention score for the w-th token with the v-th differential feature:

$$\mathbf{a}_{wv} = \frac{\exp(\mathbf{s}_{wv}/\tau_2)}{\sum_{k=1}^{V} \exp(\mathbf{s}_{wk}/\tau_2)} \tag{8}$$

Here, τ_2 is the temperature parameter. The context-aware differential representation is obtained by computing the weighted sum of differential features, with weights given by the attention scores:

$$f'_{diff_w} = \sum_{v=1}^{V} \mathbf{a}_{wv} \mathbf{F}_{diff,v} \tag{9}$$

Subsequently, we calculate the dot product between the context-aware differential representation f'_{diff_w} and the keyword feature f_k as the local similarity:

$$s_{l_i} = \frac{1}{W} \sum_{w=1}^{W} f'_{diff_w} \cdot f_k \tag{10}$$

BCE loss also employed for local contrastive learning, is defined as:

$$\mathcal{L}_{\text{local}} = -\sum_{i=1}^{B} y_i \log(\sigma(s_{l_i}/\tau_3) + (1-y_i)\log(1-\sigma(s_{l_i}/\tau_3), \tag{11}$$

Finally, the coarse-to-fine granularity contrastive loss is combined, resulting in:

$$\mathcal{L}_{\text{itc}} = \mathcal{L}_{\text{local}} + \alpha\mathcal{L}_{\text{global}} \tag{12}$$

where α is a weighting factor set to 0.5.

2.5 Difference-Guided Answer Generation

In addition to Coarse-to-Fine Grained Contrastive Learning, multi-modal pre-training tasks such as Image Text Matching (ITM) and Masked Language Modeling (MLM) [1] are used in the pre-training to align visual differences with keyword features. In the Diff-VQA fine-tuning stage, difference features serve as cross-attention queries for encoding grounded questions and decoding answers, with only autoregressive language modeling [19] used to train the Diff-MedVQA model. This two-stage training approach of alignment and generation, proven effective in many multi-modal tasks [13] for converting visual representations to textual outputs, also demonstrates its efficacy in Diff-MedVQA.

3 Experiments

3.1 Dataset

Our training and validation experiments were conducted on the MIMIC-Diff-VQA dataset [6], which is constructed based on the MIMIC-CXR dataset [9] and focuses on question-answering regarding differences in chest radiograph reports. This dataset comprises 164,324 image pairs and 700,703 questions. There are seven types of questions, including Abnormality, Presence, View, Location, Type, Level, and Difference. The distribution for training, validation, and testing is 8:1:1. Additionally, the organ and disease detectors used were pre-trained on the Chest ImaGenome [23] and VinDr-CXR datasets [15], respectively.

3.2 Implementation Details

We utilize pre-extracted anatomical visual features from FasterRCNN [20] and disease visual features from ResNet101 [5], both with dimensions of 26x1024, which are then linearly transformed to 26x768. The anatomical knowledge graph's adjacency matrix is sized at 62x62, encompassing 24 anatomical parts and 38 disease abnormalities. MCCFormer [18] is composed of 2 layers with 4-head Multi-Head Self-Attention, and both the Text Encoder and Decoder are initialized with PubMedBERT. During the pre-training phase, the loss function includes Coarse-to-Fine Grained Contrastive Learning, Image-Text Matching (ITM), and

Table 1. Experimental results in MIMIC-Diff-VQA: MCCFormer, IDCPCL, and EKAID results as reported in [6]. The baseline represents our implementation using pre-extracted anatomical and disease visual features with MCCFormer.

Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE_L	CIDEr	METEOR
MCCFormer [18]	0.214	0.190	0.170	0.153	0.340	0.000	0.319
IDCPCL [24]	0.614	0.541	0.474	0.414	0.582	0.703	0.303
EKAID [6]	0.626	0.541	0.477	**0.422**	0.645	1.911	0.340
Baseline w/o pre-train	0.523	0.453	0.400	0.350	0.488	0.572	0.287
Ours	**0.630**	**0.543**	**0.479**	**0.422**	**0.662**	**2.022**	**0.403**

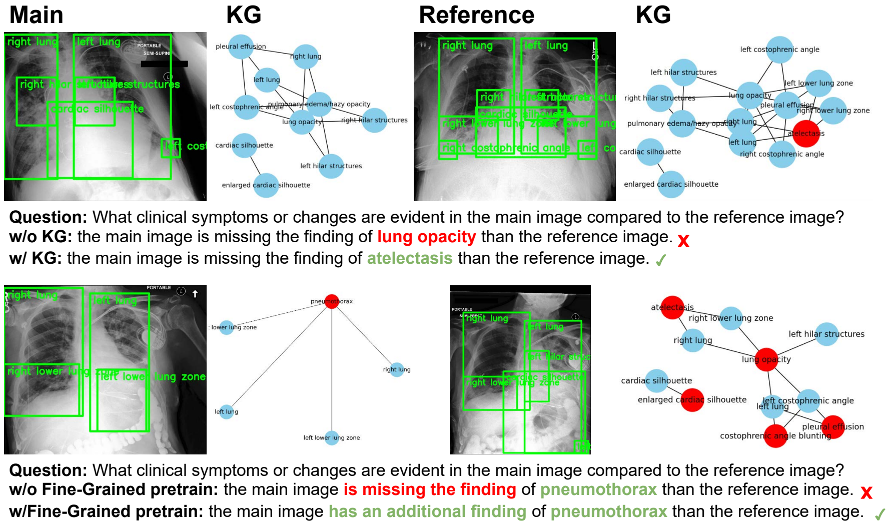
Masked Language Modeling (MLM). The batch size is set to 512. An Adam optimizer with an initial learning rate of 6e-6 and a weight decay of 1e-5 is used, and the learning rate is halved every 5 epochs. During Diff-VQA fine-tuning, we solely utilize the LM loss, maintaining a batch size of 512 consistent with pre-training, and employ an Adam optimizer with an initial learning rate of 1e-6 and a weight decay of 1e-5. Answers are generated using beam search with a beam size of 3, and evaluation metrics include CIDEr, BLEU, METEOR, and ROUGE_L [16].

3.3 Comparison with the State-of-the-Arts

Based on pre-extracted anatomical and disease visual features, we construct our baseline using MCCFormer without the anatomical knowledge graph and without pre-training, directly generating answers using Language Modeling. The results are shown in Table 1. Our best results exceed the baseline with 10.7% in BLEU1, 17.4% in ROUGE_L, and 11.6% in METEOR, also surpassing the previous top method EKAID [6] in all metrics.

Table 2. Ablation experiments in MIMIC-Diff-VQA.

Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE_L	CIDEr	METEOR
Baseline w/o pretrain							
Anatomical feat.	0.440	0.369	0.310	0.268	0.355	0.182	0.246
Disease feat.	0.459	0.385	0.333	0.287	0.377	0.203	0.254
Anatomical + Disease feat.	0.523	0.453	0.400	0.350	0.488	0.572	0.287
+ Anatomical KG	0.544	0.468	0.411	0.363	0.568	1.633	0.289
Baseline w/ pretrain							
Pretrained w/o Fine-Grained Loss	0.566	0.496	0.440	0.394	0.607	1.807	0.378
Pretrained w/ LM loss	0.592	0.518	0.460	0.411	0.660	2.010	0.392
Pretrained w/ MLM loss	**0.630**	**0.543**	**0.479**	**0.422**	**0.662**	**2.022**	**0.403**
Finetune w/ MLM loss	0.625	0.543	0.468	0.403	0.646	1.915	0.336



Question: What clinical symptoms or changes are evident in the main image compared to the reference image?
w/o KG: the main image is missing the finding of **lung opacity** than the reference image. ✗
w/ KG: the main image is missing the finding of **atelectasis** than the reference image. ✓

Question: What clinical symptoms or changes are evident in the main image compared to the reference image?
w/o Fine-Grained pretrain: the main image **is missing the finding** of **pneumothorax** than the reference image. ✗
w/Fine-Grained pretrain: the main image **has an additional finding** of **pneumothorax** than the reference image. ✓

Fig. 2. Case study of our method on MIMIC-Diff-VQA: ✓ and × denote the correctness of answers. Left and right panels show the input main and reference images, anatomical detection results, and anatomical knowledge graph, respectively.

3.4 Ablation Analysis

To demonstrate the effectiveness of our overall method, we primarily discuss the following questions: 1) The contribution of anatomical features, disease features, and the anatomical knowledge graph to performance improvement. 2) The necessity of pre-training and the impact of fine-grained contrastive learning on performance. 3) Which text generation method contributes more to generating fluent answer. The results of ablation experiments are shown in Table 2. It is evident that both anatomical and disease features benefit Diff-MedVQA, with anatomical knowledge graph providing a 2.1% and 8% increase in BLEU1 and ROUGE_L respectively. Pre-training with modal alignment of keywords substantially improves performance, especially fine-grained contrastive learning which raises BLEU1 by 6.5%. Considering the minor share of keywords in answer texts, pre-training the text encoder with Masked Language Modeling seems more beneficial than Language Modeling. Figure 2 also visualizes the proposed modules, demonstrating the performance improvement of Diff-MedVQA due to the anatomical knowledge graph and fine-grained pre-training.

4 Conclusion

In this paper, we introduce a two-stage framework for Diff-MedVQA, which acquires angle-robust visual features through an Anatomical Knowledge Graph and utilizes a Multi-Change Capture transformer to identify visual differences. The approach involves Coarse-to-Fine Granularity Contrastive Learning for aligning differential representations with keyword text, leading to answer generation via language modeling. Extensive experiments on the MIMIC-CXR-Diff dataset demonstrate the effectiveness of our proposed method.

Acknowledgments. This work was supported by the National Science and Technology Major Project under Grant 2022ZD0117103, the National Natural Science Foundation of China under Grant 62072354, the Fundamental Research Funds for the Central Universities under Grants QTZX23084 and XJSJ24015, the Natural Science Basic Research Program of Shaanxi under Grants 2024JC-YBQN-0715, 2024GX-YBXM-039 and 2024GX-ZDCYL-02-15, and the China Scholarship Council.

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

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019), https://api.semanticscholar.org/CorpusID:52967399
2. Eslami, S., Meinel, C., de Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings (2023), https://api.semanticscholar.org/CorpusID:258378275
3. Farhan, A.M.Q., Yang, S.: Automatic lung disease classification from the chest x-ray images using hybrid deep learning algorithm. Multimedia Tools and Applications pp. 1 – 27 (2023), https://api.semanticscholar.org/CorpusID:257690179
4. Gu, Y., Tinn, R., Cheng, H., Lucas, M.R., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**, 1 – 23 (2020), https://api.semanticscholar.org/CorpusID:220919723
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), https://api.semanticscholar.org/CorpusID:206594692
6. Hu, X., Gu, L., An, Q.A., 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. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2023), https://api.semanticscholar.org/CorpusID:260125237
7. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19809–19818 (2023), https://api.semanticscholar.org/CorpusID:259203473
8. Irvin, J.A., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R.L., Shpanskaya, K.S., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C., Patel, B.N., Lungren, M.P., Ng, A.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI Conference on Artificial Intelligence (2019), https://api.semanticscholar.org/CorpusID:58981871

9. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6** (2019), https://api.semanticscholar.org/CorpusID:209342303
10. Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. ArXiv **abs/1609.02907** (2016), https://api.semanticscholar.org/CorpusID:3144218
11. Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: A survey on biomedical image captioning. ArXiv **abs/1905.13302** (2019), https://api.semanticscholar.org/CorpusID:173188095
12. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: Descriptor : A dataset of clinically generated visual questions and answers about radiology images (2018), https://api.semanticscholar.org/CorpusID:53712941
13. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022), https://api.semanticscholar.org/CorpusID:246411402
14. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3334–3343 (2023), https://api.semanticscholar.org/CorpusID:257631847
15. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H., 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., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M.S., Vu, V.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9** (2020), https://api.semanticscholar.org/CorpusID:229923729
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (2002), https://api.semanticscholar.org/CorpusID:11080756
17. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4623–4632 (2019), https://api.semanticscholar.org/CorpusID:119189647
18. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1951–1960 (2021), https://api.semanticscholar.org/CorpusID:232380403
19. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
20. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 1137–1149 (2015), https://api.semanticscholar.org/CorpusID:10328909
21. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7433–7442 (2023), https://api.semanticscholar.org/CorpusID:258179419
22. Tu, Y., Li, L., Su, L., Zha, Z.J., Yan, C., Huang, Q.: Self-supervised cross-view representation reconstruction for change captioning. 2023 IEEE/CVF Interna-

tional Conference on Computer Vision (ICCV) pp. 2793–2803 (2023), `https://api.semanticscholar.org/CorpusID:263136677`

23. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest imagenome dataset for clinical reasoning. ArXiv **abs/2108.00316** (2021), `https://api.semanticscholar.org/CorpusID:235420881`
24. Yao, L., Wang, W., Jin, Q.: Image difference captioning with pre-training and contrastive learning. In: AAAI Conference on Artificial Intelligence (2022), `https://api.semanticscholar.org/CorpusID:246680391`
25. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A.L., Xu, D.: When radiology report generation meets knowledge graph. In: AAAI Conference on Artificial Intelligence (2020), `https://api.semanticscholar.org/CorpusID:211171529`