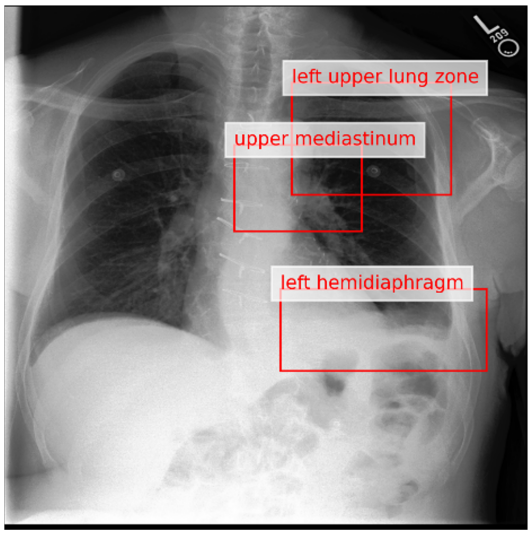
1 VISUALIZATIONS

In this section, we present visualizations of our proposed model along with other baseline models that demonstrate our superior interpretability. Firstly, we compare the visualization results of non-"Difference" questions with MMQ. As shown in Fig. 2, MMQ accurately predicted the answer. However, the GradCAM heatmap was not helpful in providing valuable information on the visualization of Regions of Interest (ROIs) as it focused on the radiographic marker on the upper right side of the image that is irrelevant to the diagnosis. On the other hand, our method not only provided the correct answer but also highlighted the pertinent region on the image, leading to the correct diagnosis.

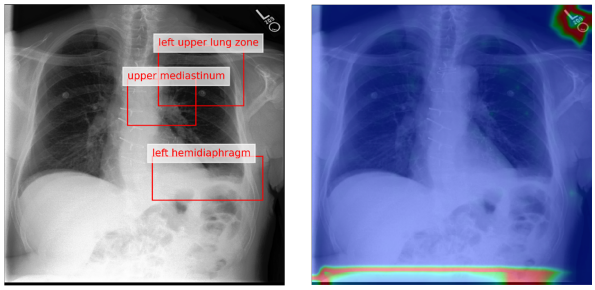


Question: where in the image is the pleural effusion located?

GT answer: left side

Prediction: left side

Figure 1: Comparison of ROIs Visualization between MMQ and Our Method for Non-"Difference" Questions.



(a) Ours

Question: where in the image is the pleural effusion located?

GT answer: left side

Prediction: left side

(b) MMQ

Question: where in the image is the pleural effusion located?

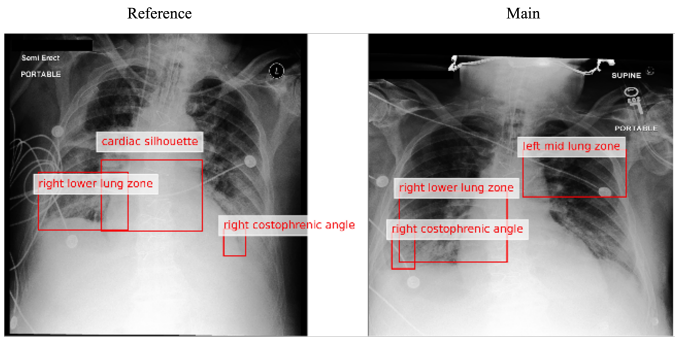
GT answer: left side

Prediction: left side

Figure 2: Comparison of ROIs Visualization between MMQ and Our Method for Non-"Difference" Questions.

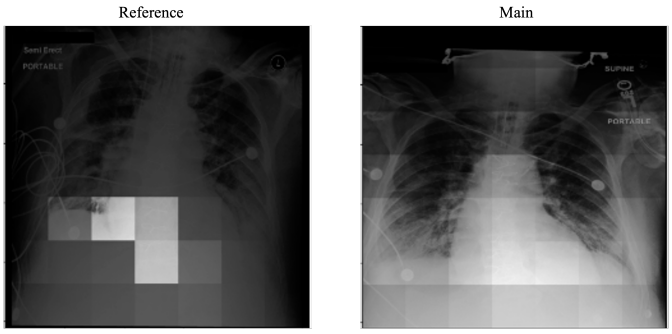
When analyzing "Difference" questions, we compare the visualization of our proposed method with all three "difference captioning" methods. We asked about changes in the right lung area, both our model and the ground truth (GT) answer indicate that the level of pleural effusion has decreased from moderate to small, as shown in Fig. 3. However, IDCPCL and MCCFormers' answers mention lung opacity, cardiomegaly, and edema in an unrelated manner, as shown in Fig. 4 and Fig. 5. In the visualized results, our model highlights the relevant regions in the right lung area, indicating that more right lung regions are selected in the main image, which is consistent with the GT answer

and the question. In contrast, IDCPCL only selects the region closest to the heart and generates irrelevant answers. Similarly, MCCFormers only selects an unrelated stripe of regions in the image, and the region of interest in the right lung, which is most relevant to the question, was not even selected.



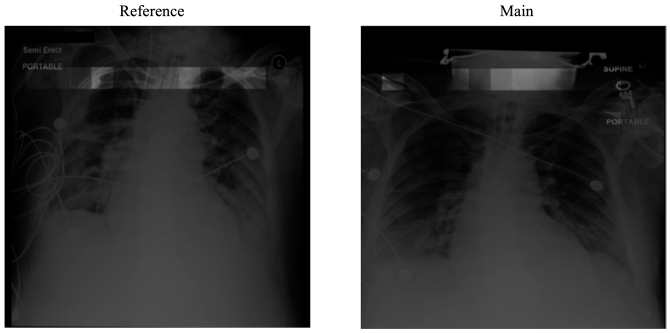
Question: what has changed in the right lung area?
GT answer: the level of pleural effusion has changed from moderate to small.
Prediction: the level of pleural effusion has changed from moderate to small.

Figure 3: ROIs Visualization of our method for "Difference" questions.



Question: what has changed in the right lung area?
GT answer: the level of pleural effusion has changed from moderate to small.
Prediction: the main image has additional findings of lung opacity , and cardiomegaly than the reference image . the main image is missing the findings of cardiomegaly , and than the the reference image . .

Figure 4: ROIs Visualization of IDCPCL for "Difference" questions.



Question: what has changed in the right lung area?
GT answer: the level of pleural effusion has changed from moderate to small.
Prediction: the main image has additional findings of pleural effusion , and edema than the reference image . the main image is missing the findings of lung opacity, and pneumonia than the reference image . the main image is missing the finding of pneumonia than the reference image …

Figure 5: ROIs Visualization of MCCFormers on "Difference" question.

2 RELATED WORK

2.1 Image Difference Captioning.

The analysis of different images has been explored by a number of researchers in the general domain. The exploration of Image Difference Captioning(IDC) can be split into three phases. The beginning of the first phase is characterized by the "spot-the-diff" dataset [9], which consists of different frames of the same video surveillance footage. This marks the first time that the IDC task has been proposed. In this phase, the researchers only focus on the pixel-level difference in the same view of the same scene. [9] use the clusters of differing pixels as a proxy for exposing object-level differences. [17, 25] propose to employ encoder-decoder architecture with attention modules to find the relationship between two images. In the second phase, the challenge was upgraded by adding different view angles of the scenes. This demands a higher requirement for the analysis of different regions between images. The iconic dataset in this phase is the CLEVR-change dataset [18], which comprises pictures of a group of objects(cube, sphere, and cylinder) from different views. The attention mechanism is widely employed to address this challenge [10, 18, 19, 23, 24, 26]. [8] propose to use an auxiliary task to enhance the primary task to generate the captions. [14] consider 3D information and adopt a scene graph to assist in localizing the changing objects. [10] also introduces a CLEVR-DC dataset, which is similar to CLEVR-change, but with a larger viewpoint change. The image pairs show more fine-grained visual differences in the third phase. The Birds-to-Words dataset [4] is composed of a variety of bird images, and each image pair is captioned by human observers. Since the species, posture, and background of the birds in each picture vary greatly, this desires a new method to solve the problem. [4] proposed Neural Naturalist, a transformer-based model. [27] learns to understand the semantic structures while comparing the images by leveraging image segmentation with a novel semantic pooling and using graph convolutional networks to perform reasoning. [30] embrace the pre-training technique to align the visual difference and the text descriptions and achieve state-of-the-art performance. We compared our method with theirs and outperformed them on our medical image difference dataset.

2.2 Medical Visual Question Answering.

Medical visual question answering aims to answer clinical questions given medical images. Medical images span a wide spectrum of modalities, including CT/MRI imaging, histopathology images, angiography, characteristic imaging appearance, ultrasound, and radiographs [2, 7, 12]. Clinical questions mainly ask for modality, plane, organ system, and abnormality [2]. However, large and well-annotated medical VQA datasets are still in scarcity. Previous MED-VQA methods mostly employ a two-stage procedure: 1) extract visual features on medical images through a detection model like Faster-RCNN [21], YOLO [20], and extract question features via BERT [3]; 2) attempt to aggregate visual and question features for predicting the final answer [1, 22, 28, 32, 34]. [12] deploys existing VQA models, i.e., the stacked attention network (SAN) [29] and the multimodal compact bilinear pooling (MCB) [5], in general domains to solve MED-VQA. [16] proposes to mix enhanced visual features framework with different attention mechanisms such as bilinear attention network (BAN) [11] and SAN. [32] proposes separate reasoning modules for different questions to improve the reasoning on medical questions. [22] integrates question categories and question topic distributions to assist answer prediction. [28] improves the CNN feature extractor with global average pooling to boost classification. [34] applies some image enhancement methods by reconstructing with small random rotations, offsets, scaling, and clipping to boost classification. However, the MED-VQA problem still lacks fine-grained annotations on images, massive diversity of medical data types, and medical reasoning skills from professions, and is thus far from practical.

2.3 Other related work.

In the general domain, NS-VQA [31] proposed to extract regions of interest(ROIs) with predicted semantic labels and generate scene graphs based on the semantic labels using Mask-RCNN. However, NS-VQA focuses on leveraging pre-designed Python logical programs to process different questions and interpret(calculate) the answers. Its answer generation greatly relies on the quality of the object segmentation and labeling by pre-trained Mask-RCNN. Since NS-VQA only evaluated the performance on a simple dataset: CLVER, where all pictures have a single color background, each object has a fixed number of labels and the same label types. Thus, training Mask-RCNN to detect different objects on this dataset is easy to obtain an ideal performance.

[15] proposed to extract abnormality-related image features by constructing a pool of normal chest X-ray images and using contrastive learning to distill the contrastive features between abnormal and normal images to improve the report generation performance. However, We focus on comparing the past and current visiting images from the same patient to track the subtle changes between the two visits. Our method is clinically driven and aims at helping the radiologist validate the hypothesis of what has changed after the intervention for each patient.

3 MORE DETAILS REGARDING DATASET AND METHOD

3.1 MIMIC-Diff-VQA construction

Tab. 1 presents the list of disease keywords, while Tab. 2 lists the attribute keywords. Tab. 3 provides a full list of questions for each question type.

Table 1: Applicable disease names

id	Disease names
0	pleural effusion
1	atelectasis
2	cardiomegaly
3	enlargement of the cardiac silhouette
4	edema
5	hernia
6	vascular congestion
7	hilar congestion
8	pneumothorax
9	heart failure
10	lung opacity
11	pneumonia
12	tortuosity of the descending aorta
13	scoliosis
14	gastric distention
15	hypoxemia
16	hypertensive heart disease
17	hematoma
18	tortuosity of the thoracic aorta
19	contusion
20	emphysema
21	granuloma
22	calcification
23	pleural thickening
24	thymoma
25	blunting of the costophrenic angle
26	consolidation
27	fracture
28	pneumomediastinum
29	air collection

3.2 Anatomical structure detection

The Anatomical structure detection results are shown in Tab. 4. Our model heavily relies on anatomical detection results, therefore, we conducted tests using our trained Faster-RCNN on both diffuse and non-diffuse diseases to assess the robustness of our detection system.

Table 2: Attribute keywords for level, location(pre), location(post), and type.

Attribute			
level	location(pre)	location(post)	type
moderate	mid to lower	the lower lobe	interstitial
acute	left	the upper lobe	layering
mild	right	the middle lobe	dense
small	retrocardiac	the left lung base	parenchymal
moderately	pericardial	the right lung base	compressive
severe	bibasilar	the lung bases	obstructive
moderate to large	bilateral	the left base	linear
moderate to severe	basilar	the right base	plate-like
mild to moderate	apicolateral	the right upper lung	patchy
moderate to large	basal	the left upper lung	ground-glass
minimal	left-sided	the right middle lung	calcified
mildly	lobe	the left middle lung	scattered
subtle	lung	the right mid lung	interstitial
massive	area	the left mid lung	focal
minimally	right-sided	the right lower lung	multifocal
increasing	apical	the left lower lung	multi-focal
decreasing	pleural	the right upper lobe	loculated
minor	upper	the left upper lobe	hazy
trace	lower	the right middle lobe	
	middle	the left middle lobe	
	mid	the right mid lobe	
	rib	the left mid lobe	
		the right lower lobe	
		the left lower lobe	
		the left apical area	
		the left apical region	
		the right apical area	
		the right apical region	
		the apical region	
		the apical area	
		the right mid to lower lung	
		the left mid to lower lung	
		the medial right lung base	
		the medial left lung base	
		the upper lungs	
		the lower lungs	
		the upper lobes	
		the lower lobes	
		the right mid to lower hemithorax	
		the soft tissues	
		the right midlung	
		the left midlung	

We select interstitial edema as the diffuse disease. Diffuse diseases accounted for 5 out of a total of 200 examples, and non-diffuse diseases accounted for 195 examples.

3.3 Relation-Aware Graph Attention Network.

As shown in Fig.6, we construct the multi-relationship graph for both main and reference images and use the relation-aware graph attention network (ReGAT) proposed by [13] to learn the graph representation for each image, and embed the image into the final latent feature. In a relation-aware graph attention network, edge labels are embedded to calculate the attention weights between nodes. Please refer to Appendix. 3.5 for details of the calculation. For simplicity, we use $G_{spa}(\cdot)$, $G_{sem}(\cdot)$, and $G_{imp}(\cdot)$ to represent the spatial graph module, the

Table 3: Full list of examples for each question type.

Question type	example
Abnormality	what abnormalities are seen in this image? what abnormalities are seen in the *[location]*? is there evidence of any abnormalities in this image? is this image normal?
Presence	is there evidence of *[abnormality]* in this image? is there *[abnormality]*? is there *[abnormality]* in the *[location]*?
View	which view is this image taken? is this PA view? is this AP view?
Location	where in the image is the *[abnormality]* located? where is the *[abnormality]*? is the *[abnormality]* located on the left side or right side? is the *[abnormality]* in the *[location]*?
Level	what level is the *[abnormality]*?
Type	what type is the *[abnormality]*?
Difference	what has changed compared to the reference image? what has changed in the *[location]* area?

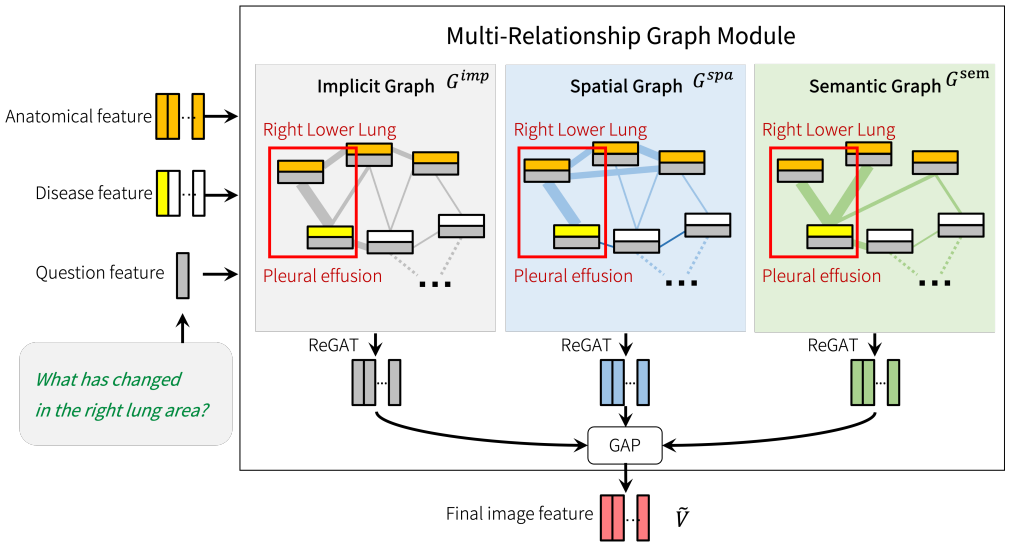


Figure 6: Multi-modal relationship graph module.

semantic graph module, and the implicit graph module, respectively. Given the input feature nodes $\mathbf{V}$ of each image, the final graph feature $\widetilde{V}$ can be represented as:

$$\widetilde{\mathbf{V}} = GAP(G_{spa}(\mathbf{V}) + G_{sem}(\mathbf{V}) + G_{imp}(\mathbf{V})) \tag{1}$$

where $GAP(\cdot)$ means the global average pooling. The image difference graph features $\widetilde{\mathbf{V}}^{diff}$ is constructed by subtracting the node feature and edge feature between the main and reference image:

$$\widetilde{\mathbf{v}}_i^{diff} = \widetilde{\mathbf{v}}_i^{main} - \widetilde{\mathbf{v}}_i^{ref}, i = 1, \cdots, 2N, \tag{2}$$

where $\widetilde{\mathbf{v}}_i^{diff}, \widetilde{\mathbf{v}}_i^{main}, \widetilde{\mathbf{v}}_i^{ref} \in \mathbb{R}^d$ represent the final feature for the i-th node of graphs. Therefore, the final graph features $\widetilde{\mathbf{V}}^{diff}, \widetilde{\mathbf{V}}^{main}$, and $\widetilde{\mathbf{V}}^{ref} \in \mathbb{R}^{2N \times d}$ can be obtained.

Table 4: Anatomical structure detection results. Precision represents when the Intersection over Union(IoU) threshold is set to 0.5.

Category	Precision (IoU =0.5)	Diffuse disease Precision	Non-diffuse Precision
right lung	97.561	100	97.569
right lower lung zone	88.774	100	88.72
right costophrenic angle	68.294	80.198	68.178
left upper lung zone	95.075	100	95.114
left hilar structures	90.092	100	90.479
left hemidiaphragm	76.314	72.277	76.908
left clavicle	83.859	100	83.808
svc	87.734	100	87.729
right atrium	80.54	100	80.457
right upper lung zone	95.55	100	95.562
right hilar structures	92.887	100	92.877
right hemidiaphragm	83.766	100	83.7
left mid lung zone	87.251	100	87.774
left apical zone	92.654	100	93.312
trachea	89.421	100	89.444
aortic arch	90.951	100	90.957
cardiac silhouette	90.643	100	90.812
carina	45.423	30.693	45.821
right mid lung zone	91.776	100	91.754
right apical zone	93.352	100	93.354
left lung	96.695	100	96.942
left lower lung zone	82.534	100	83.01
left costophrenic angle	63.95	80.198	64.321
right clavicle	87.384	100	87.393
upper mediastinum	95.216	100	95.26
cavoatrial junction	66.503	100	65.747

3.4 Feature Attention and Answer Generation

Following previous work [26], the generated main, reference, and difference features $\widetilde{\mathbf{v}}_i^{main}, \widetilde{\mathbf{v}}_i^{ref}, \widetilde{\mathbf{v}}_i^{diff}$ are applied with feature attention, then output the final feature vectors $\mathbf{l}_m$, $\mathbf{l}_r$, and $\mathbf{l}_{diff}$. For details of the calculation, please refer to Appendix 3.6. Next, we use an Answer Generation module that is composed of LSTM networks and attention modules, similar to [26]'s setting, to generate the final answer. For the calculation details, please also refer to Appendix 3.7. We use a generative language model because our questions often have a wide range of potential answers (e.g. the *difference* type question). Due to the complicated disease relationships, our dataset has a large and varied pool of answer candidates (51040 answers). Training a simple one-hot encoding classification model for these complicated questions and answers is not practical. Language models, however, can capture the semantic relationship between questions and complicated answers to generate semantic meaningful answers.

3.5 Relation-Aware Graph Attention Network

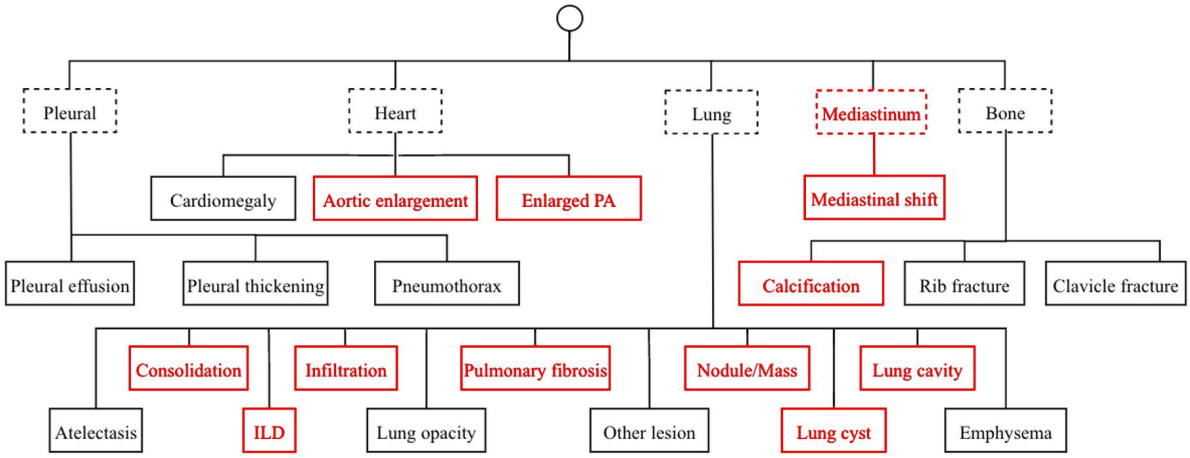
For the implicit relationship, each updated node $\widetilde{\mathbf{v}}_i \in \mathbb{R}^d$ in the final graph can be calculated as below:

$$\widetilde{\mathbf{v}}_i = \mathbf{W}^o \cdot (\|_{m=1}^{M} \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}^m \mathbf{v}_j)) \tag{3}$$

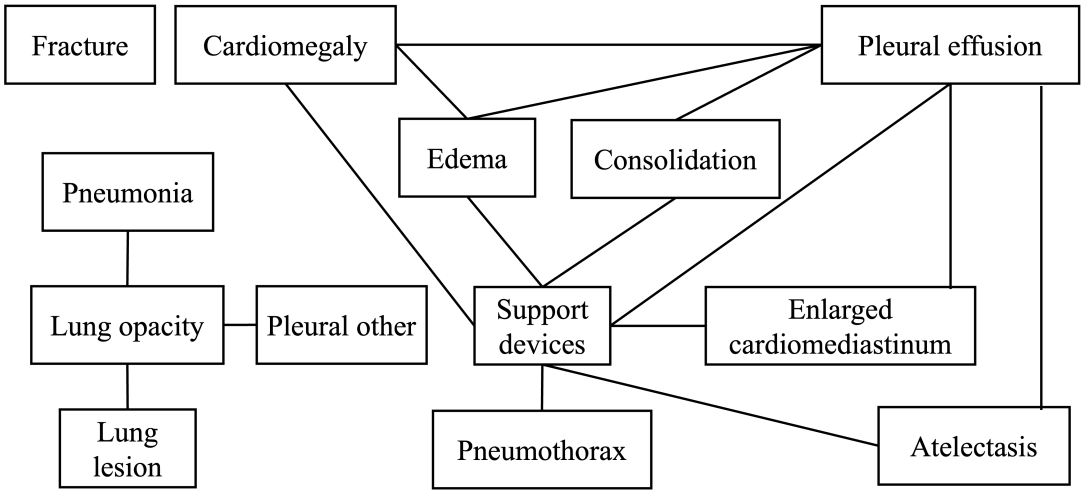
where $\mathcal{N}_i$ is the neighborhood set of the node i, $\mathbf{W}^m \in \mathbb{R}^{d \times (d_f + d_q)}$ is the projection matrix, d is the dimension of the final node feature, σ is the activation function, $\|_{m=1}^{M}$ represents concatenating the output of the M attention heads, $\mathbf{W}^o \in \mathbb{R}^{d \times Md}$. The attention weights α_{ij} between the node i and node j consider the similarity between node pairs and the relations between the corresponding region locations. The calculation for α_{ij} can be formulated as:

$$\alpha_{ij} = \frac{\alpha_{ij}^b \cdot \exp(\alpha_{ij}^v)}{\sum_{j=1}^{K} \alpha_{ij}^b \cdot \exp(\alpha_{ij}^v)} \tag{4}$$

$$\alpha_{ij}^v = (\mathbf{U}\mathbf{v}_i)^\top \cdot (\mathbf{V}\mathbf{v}_j) \tag{5}$$



(a) Anatomical knowledge graph (compared to [33], our new added disease types are annotated by red.)



(b) Co-occurance knowledge graph

Figure 7: Knowledge graphs

$$\alpha_{ij}^{b} = \max(0, w \cdot f_b(\mathbf{b}_{ij})) \tag{6}$$

where $U, V \in \mathbb{R}^{d\times(d_f+d_q)}$ are projection matrices. $\mathbf{b}_{ij}$ is the relative geometry feature between node i and j, and can be calculated by $[\log(\frac{|x_i-x_j|}{w_i}), \log(\frac{|y_i-y_j|}{h_i}), \log(\frac{w_j}{w_i}), \log(\frac{h_j}{h_i})]$, f_b is a function that embeds the 4-dimensional relative geometry feature into d-dimensional,$w \in \mathbb{R}^d$ is a vector that transforms the feature into a scalar weight. The bounding box coordinates, widths, and heights of the node i and j can be represented by $x_i, x_j, y_i, y_j, w_i, w_j, h_i$, and h_j.

Spatial and semantic graphs, which can also be called explicit graphs, can be seen as directed graphs. The updating rule considers the relation directions between node pairs and the labels of the edges. The formulation of a single attention head is shown below:

$$\widetilde{\mathbf{v}}_i = \sigma\Big(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\mathbf{W}_{dir(i,j)}\mathbf{v}_j + b_{lab(i,j)}\Big) \tag{7}$$

$$\alpha_{ij} = \frac{\exp\left((\mathbf{U}\mathbf{v}_i)^\top \cdot \mathbf{V}_{dir(i,j)}\mathbf{v}_j + c_{lab(i,j)}\right)}{\sum_{j\in\mathcal{N}_i}\exp\left((\mathbf{U}\mathbf{v}_i)^\top \cdot \mathbf{V}_{dir(i,j)}\mathbf{v}_j + c_{lab(i,j)}\right)} \tag{8}$$

where $dir(i, j)$ represents the direction goes from node i to j, $lab(i, j)$ is the label assigned to the edge (i, j), $W_{dir(i,j)}, V_{dir(i,j)} \in \mathbb{R}^{d\times(d_f+d_q)}$ are projection matrices, $b_{lab(i,j)}, c_{lab(i,j)} \in \mathbb{R}^d$ are bias terms. The multi-head attention can be calculated similarly by concatenating the output features and adding a projection matrix $\mathbf{W}^o \in \mathbb{R}^{d\times Md}$.

3.6 Feature Attention Module

The generated main image features $\widetilde{\mathbf{V}}_i^{main}$, reference image feature$\widetilde{\mathbf{V}}_i^{ref}$ and the difference feature$\widetilde{\mathbf{V}}_i^{diff}$ are then fed into the Feature Attention Module, which is similar to the two modules in [26] called Cross-semantic Relation Measuring block(CSRM) and Prior Knowledge-guided Change Localizer. In the Feature Attention module, we first calculate the prior knowledge C'_m, and C'_r for the main image and the reference image, respectively. Take C'_m for example, the calculation process is shown below.

$$C_m = \phi(\widetilde{\mathbf{V}}^{main} W_q^c + \widetilde{\mathbf{V}}^{main} W_v^c + b^c) \tag{9}$$

$$A_m = \sigma(\widetilde{\mathbf{V}}^{main} W_q^a + \widetilde{\mathbf{V}}^{main} W_v^a + b^a) \tag{10}$$

$$C'_m = A_m \odot C_m \tag{11}$$

where $C_m \in \mathbb{R}^{2N \times d}$ is the "candidate change", $A_m \in \mathbb{R}^{2N \times d}$ is the "attention gate", $W_q^c, W_v^c, W_q^a, W_v^a \in \mathbb{R}^{d \times d}$, $b^c, b^a \in \mathbb{R}^d$, $\odot$ represents the element-wise multiplication, ϕ is the tanh function, σ is the sigmoid function. C'_r can be calculated similarly.

Then, guided by prior knowledge, we calculate the attention weights a_m and a_r for the main image and the reference image, respectively. The formulations are shown below:

$$a_m = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1([\widetilde{\mathbf{V}}^{main}; \widetilde{\mathbf{V}}^{diff}; C'_m])))) \tag{12}$$

$$a_r = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1([\widetilde{\mathbf{V}}^{ref}; \widetilde{\mathbf{V}}^{diff}; C'_r])))) \tag{13}$$

where $[;]$ represents the concatenation, FC represents fully-connected layer, σ represents the sigmoid function.

After obtaining the attention weights $a_m \in \mathbb{R}^{2N}$ and $a_r \in \mathbb{R}^{2N}$, the final image feature vector $\mathbf{l}_m$ and $\mathbf{l}_r$ for the main image and the reference image can be calculated as follows:

$$\mathbf{l}_m = \sum_{i=1}^{2N} a_{m_i} \widetilde{\mathbf{v}}_i^{main} \tag{14}$$

$$\mathbf{l}_r = \sum_{i=1}^{2N} a_{r_i} \widetilde{\mathbf{v}}_i^{ref} \tag{15}$$

where $a_m \in \mathbb{R}^{2N}$ and $a_r \in \mathbb{R}^{2N}$ are the attention weights. The difference vector is accordingly computed as:

$$\mathbf{l}_{diff} = \mathbf{l}_m - \mathbf{l}_r \tag{16}$$

3.7 Answer Generation

Dynamic Feature Generation. At each time step t, we first calculate the attention weights $\alpha_i^{(t)}$, which is for calculating the intermediate dynamic feature $l_{dyn}^{(t)}$ in the next step. The $\alpha_i(t)$ can be calculated as follows:

$$v = \mathrm{ReLU}(W_{a_1}[l_{bef}; l_{diff}; l_{aft}] + b_{a_1}) \tag{17}$$

$$u^{(t)} = [v; h_c^{(t-1)}] \tag{18}$$

$$h_a^{(t)} = LSTM_a(h_a^{(t)} | u^{(t)}, h_a^{(0:t-1)}) \tag{19}$$

$$\alpha_i^{(t)} \sim Softmax(W_{a_2} h_a^{(t)} + b_{a_2}) \tag{20}$$

where $W_{a_1}, W_{a_2}, b_{a_1}, b_{a_2}$ are learnable parameters, $LSTM_a$ is a LSTM network used as attention weights generator, $h_a^{(t)}$ is the output of the $LSTM_a$ at the time step t, $h_c^{(t-1)}$ is the output of the answer generator $LSTM_c$ at the time step $t-1$, which will be explained in more detail later.

Then, the intermediate dynamic feature $l_{dyn}^{(t)}$ can then be calculated as follows:

$$l_{dyn}^{(t)} = \sum_i \alpha_i^{(t)} l_i \tag{21}$$

where $i \in (bef, diff, aft)$.

Before calculating the final dynamic feature $L_{dyn}^{(t)}$, POS feature $p^{(t)}$ needs to be obtained first. The POS feature is calculated from the hidden embedding of the answer $h_c^{(t-1)}$ from the last time step. The calculation can be formulated as below:

$$h_p^{(t)} = \mathrm{ReLU}(W_{p_1} h_c^{(t-1)} + b_{p_1}) \tag{22}$$

$$w_p^{(t)} = Softmax(W_{p_2} h_p^{(t)} + b_{p_2}) \tag{23}$$

$$p^{(t)} = E_p w_p^{(t)} \tag{24}$$

where $W_{p_1}, W_{p_2}, b_{p_1}, b_{p_2}$ are learnable parameters, E_p is a learnable POS embedding matrix.

With the intermediate dynamic feature $l_{dyn}^{(t)}$ and the POS feature $p^{(t)}$, we can calculated the final dynamic feature $L_{dyn}^{(t)}$.

Table 5: Results of classification-based VQA problem.

Answer	SYSU-HCP	Ours
Pneumothorax	0.806	0.876
edema	0.737	0.893
lung lesion	0.665	0.843
no	0.537	0.951
lung opacity	0.605	0.859
ateletasis	0.645	0.868
pleural other	0.858	0.845
support devices	0.769	0.924
pneumonia	0.715	0.833
pleural effusion	0.796	0.938
enlarged cardiomediastinum	0.725	0.828
yes	0.545	0.944
consolidation	0.708	0.819
cardiomegaly	0.688	0.892
fracture	0.664	0.871
total (micro)	0.792	0.934
total (macro)	0.697	0.879

$$\beta_t = \sigma(W_{c_2}(\mathrm{ReLU}(W_{c_1}[p^{(t)}; h_c^{(t-1)}; l_{dyn}^{(t)}]))) \tag{25}$$

$$L_{dyn}^{(t)} = \beta_t \odot l_{dyn}^{(t)} \tag{26}$$

where the range of β_t is $[0, 1]$, the value of it indicates how much the visual information will be used in the answer generation part.

Answer generator. The answer is generated by an LSTM network word by word. The initial word at time step 0 is the $< start >$ token.

$$c^{(t)} = [E[w^{(t-1)}]; L_{dyn}^{(t)}] \tag{27}$$

$$h_c^{(t)} = \mathrm{LSTM}_c(h_c^{(t)}|c^{(t)}, h_c^{(0:t-1)}) \tag{28}$$

$$w^{(t)} \sim Softmax(W_c h_c^{(t)} + b_c) \tag{29}$$

where E is a word embedding layer, $E[w^{(t-1)}]$ is the word embedding for the word $w^{(t-1)}$, W_c, b_c are learnable parameters.

We adopt the generative language model because our questions have highly diverse answers. (e.g. the *difference* type question). A simple classification model is not adequate for our task.

3.8 Other results

We evaluated our proposed multi-relationship graph for the general chest X-ray image classification-based VQA problem (14 diseases) and compared it to state-of-the-art method SYSU-HCP [6], the best team in the ImageCLEF-VQA-Med 2021 task. As shown in Tab. 5, We use AUC as the metric because answering abnormality questions can be considered a multi-label classification problem. Our model achieved significant improvement compared to the state-of-the-art disease classification performance.

We show the results of our model on each question type in Tab. 6. It is worth noting that, Bleu 3 and Bleu 4 tend to have low scores. This is because the answers to most of the questions are short, except for the "Difference" questions. For abnormality questions, 72% of the answers have less than or equal to 2 words; for location questions, 79% of the answers have less than or equal to 2 words; 93% of level questions have one-word answers.

3.9 More visualization examples of our method

To prove the improvement of the interpretability of our model by adding the spatial and semantic graphs, we visualize the ROIs of our model using different graphs and demonstrate the predictions. As shown in Fig. 8(b), our model using the only implicit graph missed the regions important for the question and failed to interpret the correct answer. In contrast, as shown in Fig. 8(a), with the help of the spatial relationship graph, our model succeeded in finding the critical region and delivering the correct answer.

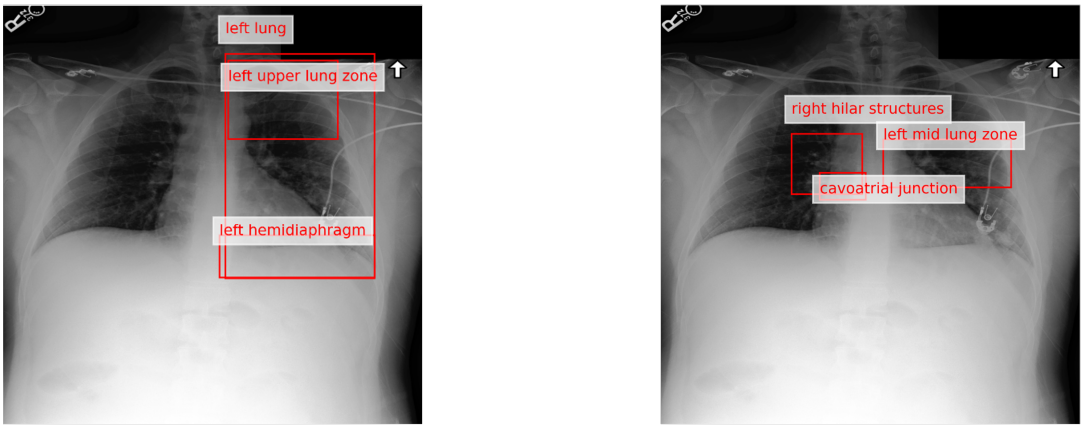
Fig. 9 demonstrates a similar scenario on an abnormality-type question. our model using only the implicit graph detected only one abnormality, atelectasis, missed pleural effusion, and lung opacity. However, with the help of the semantic relationship graph, which

Table 6: Results of each question type. "-" represents not applicable because no ground truth answer has enough words to trigger the corresponding Bleu metric.

Question type	Bleu 1	Bleu 2	Bleu 3	Bleu 4
Abnormality	0.482	0.333	0.197	0.109
Presence	0.801	-	-	-
View	0.948	0.941	-	-
Location	0.525	0.364	0.210	0.144
Level	0.496	0.101	0.068	-
Difference	0.641	0.564	0.500	0.441

emphasizes the relationship between pleural effusion, atelectasis, and lung opacity, our full model detected all three abnormalities and provided the correct answer.

Q: Is the consolidation located on the left side or right side?
GT answer: left side



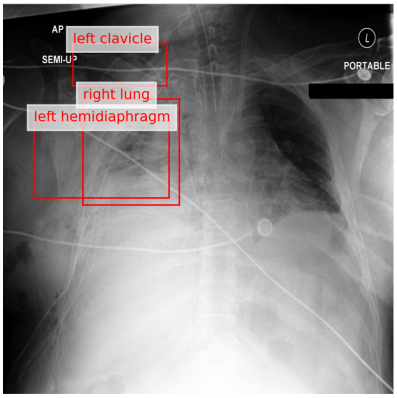
(a) Full model
Prediction: left side

(b) Model with implicit graph only
Prediction: right side

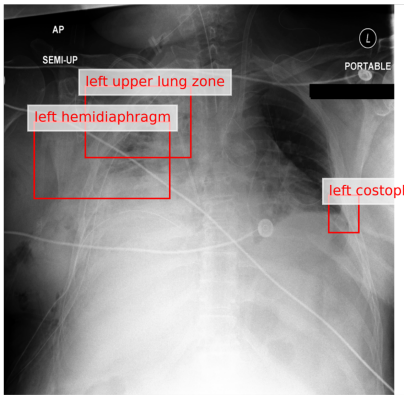
Figure 8: ROIs Visualization comparison between implicit graph and all graphs on location type question.

As shown in Fig. 10, when asking about pleural effusion, which is an abnormality that happens in the lower lung when there is excess fluid between the layers of the pleura outside the lungs, our method highlighted the corresponding regions (left lower lung). Also, by focusing on these regions, our method can accurately determine the change in the level of pleural effusion between the main and reference image. In Fig. 11, our method also highlighted cardiac silhouette, this could be because of the strong semantic relationship between cardiomegaly and pleural effusion.

Q: what abnormalities are seen in this image?
GT answer: pleural effusion, atelectasis, lung opacity



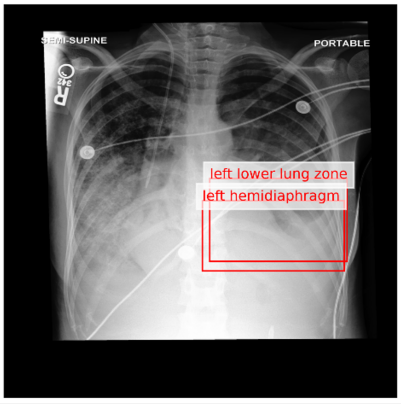
(a) Full model
Prediction: pleural effusion, atelectasis, lung opacity



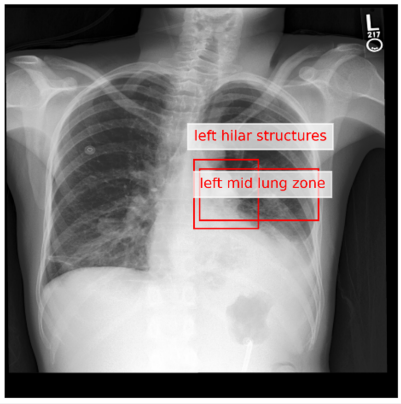
(b) Model with implicit graph only
Prediction: atelectasis

Figure 9: ROIs Visualization comparison between implicit graph and all graphs on abnormality type question.

Reference

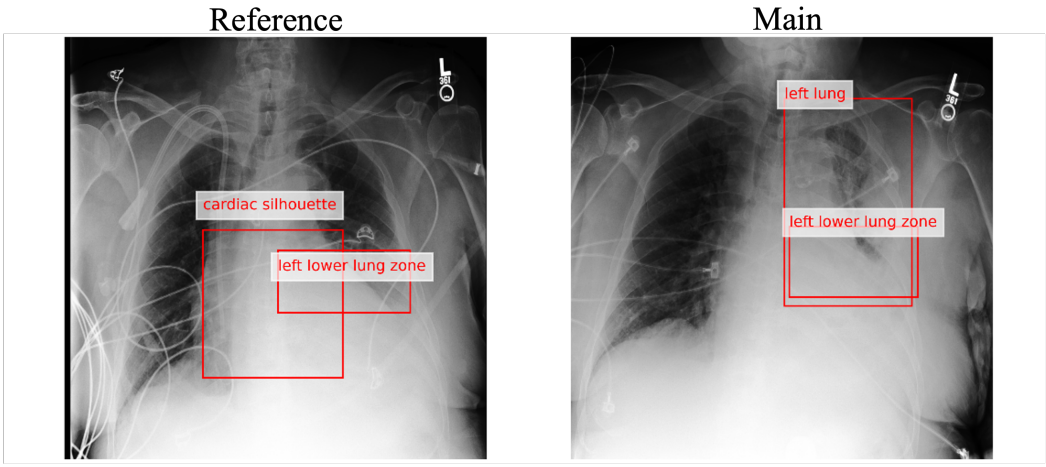


Main



Question: what has changed in the left lung area?
GT Answer: the level of pleural effusion has changed from moderate small to small.
Prediction: the level of pleural effusion has changed from moderate small to small.

Figure 10: Visualization example 1



Question: what has changed in the left lung area?

GT Answer: the level of pleural effusion has changed from moderate to small.

Prediction: the level of pleural effusion has changed from moderate to small.

Figure 11: Visualization example 2

REFERENCES

[1] Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. 2018. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain.. In *CLEF (Working Notes)*.

[2] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. *CLEF (Working Notes)* 2 (2019).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural naturalist: generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101* (2019).

[5] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[6] Haifan Gong, Ricong Huang, Guanqi Chen, and Guanbin Li. 2021. Sysu-hcp at vqa-med 2021: A data-centric model with efficient training methodology for medical visual question answering. *Proceedings http://ceur-ws. org ISSN* 1613 (2021), 0073.

[7] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).

[8] Mehrdad Hosseinzadeh and Yang Wang. 2021. Image Change Captioning by Learning from an Auxiliary Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2725–2734.

[9] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584* (2018).

[10] Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021. Agnostic Change Captioning with Cycle Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2095–2104.

[11] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems* 31 (2018).

[12] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5, 1 (2018), 1–10.

[13] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10313–10322.

[14] Zeming Liao, Qingbao Huang, Yu Liang, Mingyi Fu, Yi Cai, and Qing Li. 2021. Scene Graph with 3D Information for Change Captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5074–5082.

[15] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965* (2021).

[16] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 522–530.

[17] Ariyo Oluwasanmi, Enoch Frimpong, Muhammad Umar Aftab, Edward Y Baagyere, Zhiguang Qin, and Kifayat Ullah. 2019. Fully convolutional captionnet: Siamese difference captioning attention model. *IEEE Access* 7 (2019), 175929–175939.

[18] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4624–4633.

[19] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. 2021. Describing and Localizing Multiple Changes with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1971–1980.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[22] Lei Shi, Feifan Liu, and Max P Rosen. 2019. Deep Multimodal Learning for Medical Visual Question Answering.. In *CLEF (Working Notes)*.

[23] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *European Conference on Computer Vision*. Springer, 574–590.

[24] Yaoqi Sun, Liang Li, Tingting Yao, Tongyv Lu, Bolun Zheng, Chenggang Yan, Hua Zhang, Yongjun Bao, Guiguang Ding, and Gregory Slabaugh. 2022. Bidirectional difference locating and semantic consistency reasoning for change captioning. *International Journal of Intelligent Systems* (2022).

[25] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689* (2019).

[26] Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. 2021. Semantic Relation-aware Difference Representation Learning for Change Captioning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 63–73.

[27] An Yan, Xin Eric Wang, Tsu-Jui Fu, and William Yang Wang. 2021. L2C: Describing Visual Differences Needs Semantic Understanding of Individuals. *arXiv preprint arXiv:2102.01860* (2021).

[28] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. 2019. Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain. *CLEF (Working Notes)* 85 (2019).

[29] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[30] Linli Yao, Weiying Wang, and Qin Jin. 2022. Image Difference Captioning with Pre-training and Contrastive Learning. *arXiv preprint arXiv:2202.04298* (2022).

[31] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems* 31 (2018).

[32] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2345–2354.

[33] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12910–12917.

[34] Yangyang Zhou, Xin Kang, and Fuji Ren. 2018. Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering.. In *CLEF (Working Notes)*.