

Article

DKA-RG: Disease-Knowledge-Enhanced Fine-Grained Image–Text Alignment for Automatic Radiology Report Generation

Heng Yin ¹, Wei Wu ^{2,*}  and Yongtao Hao ^{1,*} ¹ CAD Research Center, Tongji University, Shanghai 200092, China; 2230761@tongji.edu.cn² Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China

* Correspondence: weiwu@tongji.edu.cn (W.W.); haoyt@tongji.edu.cn (Y.H.)

Abstract: Automatic radiology report generation is a task that combines artificial intelligence and medical information processing, and it fully relies on computer vision and natural language processing techniques. Nowadays, automatic radiology report generation is still a very challenging task because it requires semantically adequate alignment of data from two modalities: radiology images and text. Existing approaches tend to focus on coarse-grained alignment at the global level and do not take into account the disease characteristics of radiology images at fine-grained semantics, which results in the generated reports potentially omitting key disease diagnostic descriptions. In this work, we propose a new approach, disease-knowledge-enhanced fine-grained image–text alignment for automatic radiology report generation (DKA-RG). The method combines global and disease-level alignment, thus facilitating the extraction of fine-grained disease features by the model. Our approach also introduces a knowledge graph to inject medical domain expertise into the model. Our proposed DKA-RG consists of two training steps: the image–report alignment stage and the image-to-report generation stage. In the alignment stage, we use global contrastive learning to align images and texts from a high level and also augment disease contrastive learning with medical knowledge to enhance the disease detection capability. In the report generation stage, the report text generated from the images is more accurate in describing the disease information thanks to sufficient alignment. Through extensive quantitative and qualitative experiments on two widely used datasets, we validate the effectiveness of our DKA-RG on the task of radiology report generation. Our DKA-RG achieves superior performance on multiple types of metrics (natural language generation and clinical efficacy metrics) compared to existing methods, demonstrating that the method can improve the reliability and accuracy of automatic radiology report generation systems.

Keywords: computer vision; natural language processing; multimodal modeling; contrastive learning; radiology report generation; knowledge graph



Citation: Yin, H.; Wu, W.; Hao, Y. DKA-RG: Disease-Knowledge-Enhanced Fine-Grained Image–Text Alignment for Automatic Radiology Report Generation. *Electronics* **2024**, *13*, 3306. <https://doi.org/10.3390/electronics13163306>

Academic Editor: Hyunjin Park

Received: 22 July 2024

Revised: 11 August 2024

Accepted: 14 August 2024

Published: 20 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Medical image analysis and report writing play an important role in the field of medicine. Doctors diagnose diseases and evaluate their progression by analyzing medical images such as X-rays, CT, MRI, etc., and writing reports to develop treatment plans. Image analysis can detect early signs of diseases and provide accurate information about the pathology, thus improving the accuracy of diagnosis and the effectiveness of treatment. The report writing systematizes and standardizes the results of these analyses, making them easy for clinical doctors to refer to and use while providing patients with clear and concise health status maps. High-quality medical imaging analysis and report writing not only improve the level of medical services but also promote medical research and the application of new technologies.

Computer-based medical assistance systems have gradually begun to enter actual medical scenarios. For example, method [1] develops a computer system for asymmetry detection in mammographic images, and method [2] uses deep learning technology for

intelligent image recognition, thereby helping clinicians make more informed diagnostic decisions. Computer-assisted medical care can improve the accuracy of medical diagnosis and greatly improve the efficiency of medical diagnosis.

The writing of radiology reports is a very laborious process that requires doctors to first carefully examine and analyze radiology images, and then use professional and accurate medical terminology to record them [3]. Due to the rapid development of artificial intelligence technology, many researchers have begun to explore how these techniques can be applied to automated radiology image analysis and report generation [4–6]. Through computer vision technology, we can extract disease information from radiology images, and through natural language processing technology, we can automatically generate report text based on this information. This process can greatly reduce the workload of doctors and improve the standardization of reports, which is meaningful for promoting medical development.

Radiology report generation is very similar to general image-to-text generation tasks [7–10] but differs from them in that it pays more attention to fine-grained semantic information in radiology images and requires higher accuracy. Small changes in radiology images can correspond to different diagnostic results, so the automated generation system must be highly sensitive and precise to capture and correctly interpret these details. In addition, radiology reports must follow strict medical terminology and standards to ensure that doctors and patients can accurately understand and rely on the content. These unique requirements make report generation a complex and critical technical challenge.

Therefore, one of the difficulties in automated radiology report generation lies in achieving semantic alignment of medical images and reports, i.e., accurately establishing cross-modal correspondence between disease information at a particular site in an image and the description of the disease in a report. This requires the system to be able to accurately recognize and interpret subtle lesions in the images and to accurately represent these findings during text generation. Establishing this cross-modal correspondence is a key challenge in ensuring that the content of reports is accurate, detailed, and clinically meaningful, which is the focus of current research.

At present, some methods [11,12] utilize contrastive learning to establish semantic associations between radiology images and reports. Although they have achieved some results, such semantic associations are still limited to the global level. Some other methods perform local alignment; AlignTransformer [13] aligns image-region-level features with some fixed disease labels, and RGRG [14] aligns independent sentences with anatomical locations using manually labeled data. Although these alignment methods can align image regions to textual local descriptions at a fine-grained level, the additional annotation data make the process more costly.

In this paper, we propose a disease-knowledge-enhanced fine-grained image–text alignment for automatic radiology report generation named DKA-RG. To further facilitate the semantic alignment of radiology images and texts, our framework constructs contrastive learning methods at the global and disease levels, respectively. Specifically, we encode images and texts with Vision Transformer [15] and BERT [16] models, respectively, to obtain high-dimensional vector representations of both. Global contrastive learning is then constructed by treating matched images and reports as positive sample pairs and mismatched sample pairs as negative sample pairs. Global contrastive learning can bring semantically related image–text pair encoding closer and push away semantically unrelated image–text pairs encoding farther, thus realizing semantic alignment. The semantic alignment of radiology images and report text still has deficiencies by global alignment alone, so we use the graph attention network (GAT) [17] to obtain a series of disease tokens based on the medical disease knowledge graph. Meanwhile, we construct a disease decoder network based on Transformer [18]. We take disease tokens and the image/text features as input to the disease decoder to query the disease features in the image and text. Since there is semantically consistent disease information in the image and the corresponding report text, we construct disease-level contrastive learning. Specifically, the disease query features

of the same disease query token in the image and the text, respectively, are used as positive sample pairs, and the disease query features obtained from different disease query tokens are used as negative sample pairs, so as to promote the semantic alignment of diseases in the image and the text. With the alignment at both global and disease levels, our medical image encoder can better extract the medical features in images. Finally, we construct a Transformer-based [18] report decoder to generate the radiology report based on the image features obtained from the image encoder and disease features from the disease decoder.

Overall, our method addresses the shortcomings of existing methods in promoting medical image text alignment, achieving finer grained alignment and thus achieving better disease feature extraction capabilities. Thanks to excellent alignment of image and textual features, our method can be more reliable in practical application scenarios, further promoting the popularization of medical report automatic generation systems in clinical medicine.

Our contributions are summarized as follows:

- We build DKA-RG, a novel framework for automatic radiology report generation. The framework utilizes multi-level contrastive learning, which more comprehensively facilitates cross-modal semantic alignment of radiology images and reports during the encoding process.
- We design a disease decoder network using the medical knowledge graph, which not only takes into account medical disease relationships but also extracts disease information from images and reports at a fine-grained level. We construct disease-level alignment to improve the feature extraction capability of the image encoder.
- We conduct comprehensive qualitative and quantitative experiments on two widely used datasets, MIMIC-CXR [19] and IU-Xray [20]. The experimental results demonstrate that our method achieves high performance in terms of both language generation metrics and medical clinical evaluation metrics, outperforming existing methods. We also construct qualitative experiments to analyze the accuracy of generated reports.

2. Related Work

2.1. Contrastive Learning

Unlike normal supervised learning, contrastive learning, a self-supervised learning method, is well suited for domains that lack labeled data (such as the medical field) or where multimodal alignment is required. The basic idea of contrastive learning is to enable the model to distinguish between similar and dissimilar sample pairs, thereby enhancing the model's ability to extract data features in the process. SimCLR [21] proposed a very simple framework that utilizes data augmentation techniques to create positive and negative sample images and then trains the model using contrastive learning loss to enhance its ability to extract image features. Another work, MoCo [22], proposed momentum-based contrastive learning to address the sensitivity of contrastive learning to batch size. MoCo [22] placed encoded sample pairs into a queue during the contrastive learning process to simulate batch size expansion, thereby promoting the effectiveness of contrastive learning with limited computing resources. Contrastive learning has also made significant strides in biomedical imaging and the generation of radiology reports. ConVIRT [23] facilitates the alignment of medical images and reports using a multimodal contrastive learning approach, which improves the performance of image classification, retrieval, etc. DCL [12] introduces a dynamic knowledge graph in multi-task learning to improve report generation.

Our work utilizes contrastive learning to facilitate alignment between medical images and report text. Due to the fact that medical images and text belong to two types of modal information but have a matching relationship, they meet the requirements of contrastive learning. Firstly, we perform contrastive learning on the global features obtained from the encoder of the image and report text, which promotes global semantic alignment. In addition, we also extract disease features from images and text for contrastive learning at the disease level.

2.2. Medical Knowledge Graph

The knowledge graph can effectively represent the extensive and complex biomedical information in the medical field. By modeling the relationships and entities of medical diseases, symptoms, anatomical locations, treatment plans, etc., medical knowledge graphs can be fully utilized for disease discovery and diagnosis. Zhang et al. [5] introduces knowledge graphs into the field of radiology report generation and incorporates an attention mechanism to improve the accuracy of report generation. DCL [12] dynamizes the knowledge graph so that medical knowledge can be flexibly introduced during training and inference. KnowMat [24] combines knowledge graphs with other types of knowledge, demonstrating the importance of knowledge richness for the field of the generation of radiology reports.

These works confirm the effectiveness of medical knowledge graphs. By structuring biomedical knowledge, medical knowledge graphs can not only enhance the effectiveness of information storage, but also promote the extraction of specific disease features in medical images.

2.3. Radiology Report Generation

The automatic radiology report generation task combines computer vision and natural language processing techniques involving information from both modalities. Its purpose is to assist physicians in disease diagnosis by automatically generating detailed and accurate text descriptions based on medical images using computer technology.

Earlier approaches directly introduced generalized image caption generation methods into the medical field. Xu et al. [8] proposed a method that combines a CNN for image feature extraction and an RNN for sequence generation and enhances attention to important parts of the image through an attention mechanism. The model has inspired subsequent work in the medical field, where the attention mechanism helps guide the model to extract information in the image that is more important for diagnosis. The introduction of the encoder–decoder architecture further facilitated the technique of automatic radiology report generation. The model proposed by Jing et al. [4] uses CNNs to encode and extract medical image features and RNNs to decode image features into textual descriptions. The method introduces a special attention mechanism that focuses on the important regions of the image and text at the same time, which improves the accuracy of the generated report. Yuan et al. [25] extended this approach by incorporating a hierarchical recurrent neural network model. The model simulates the real report writing process in two steps: first creating high-level impressions and then describing the detailed findings section. Yi et al. [26] proposes two-stage global enhancement layers to facilitate the generation of more reliable reports from a global perspective. Gu et al. [27] proposed using additional mask information as guidance to promote the targetedness of medical report generation.

Some work has begun to introduce pre-trained large models to improve report quality. A number of BERT [16] and its variant models can be used to improve the quality of generated reports. Zhang et al. [28] proposed a hybrid model in which CNN is used to facilitate image feature extraction, and BERT is used to facilitate the richness and fluency of the context of the report. Pre-trained models contain a lot of prior knowledge, which can increase the richness of the generated reports.

With the rise of large language models, many researchers have begun to try to use large language models for the automatic generation of medical reports. The method [29] uses ChatGPT [30] for automatic writing of medical reports, which is verified to have excellent readability and fluency. In addition, the method [31] focuses on the key exception information in the report, thus avoiding the templating of report generation.

In addition, it is a challenging task to better integrate AI techniques into actual clinical healthcare workflows. Method [1] develops a computer system for the detection of asymmetry in mammographic images. Nakaura et al. [32] develops pipelines to automatically convert AI results into structured reports, which drives the practical application of AI-assisted healthcare.

These existing methods for building models for medical image analysis and report generation, while supporting the fusion of information from both modalities of medical images and reports, do not fully consider the fine-grained semantic alignment of medical information. Our approach improves the alignment of multimodal characteristics at the disease level, which is more adapted to the medical reporting domain. In Table 1, we present a comparison between our method and existing methods, as well as the advantages of our method.

Table 1. Summary of the advantages and disadvantages of different types of medical report automatic generation methods, as well as the advantages of our method compared to these methods.

Method Type	Representative Methods	Advantage	Disadvantage
Non Alignment	Jing et al. [4] Wang et al. [33]	Simple end-to-end training	Weak ability to extract disease features
Global Alignment	Ma et al. [34] Li et al. [12]	Unify multimodal features of image and text to enhance feature extraction capabilities	Lack of fine-grained image text feature alignment
Knowledge enhancement	Huang et al. [35] Yang et al. [24]	Medical knowledge enhances performance	The way of introducing knowledge is too complicated
The advantages of our method	Global and local alignment of diseases enhances the ability to extract disease features. The disease knowledge graph serves as a direct guide for contrastive learning, which is simple and efficient.		

3. Method

3.1. Radiology Image and Text Feature Extraction

Our architecture consists of six modules which are image encoder, text encoder, graph encoder, image disease decoder, text disease decoder, and text decoder, as shown in Figure 1.

Image Feature Extraction. With the rise of the Transformer [18] architecture, vision encoders such as Vision Transformer (ViT) [15] have powerful visual feature extraction performance. Therefore, we use it as the image encoder. The structure of ViT is shown in Figure 2. Given a medical image I , it is first preprocessed by scaling, adjusted to the size required by the model (resolution 224×224) and normalized. Next, the image I is cut into multiple small patches of 16×16 size. Each patch p_i is mapped to a high-dimensional feature vector I_i^f through a linear layer. In order for the model to process the sequence information, each vector of features I_i^f is associated with a position encoding pos_i , which is used to help the model understand the relative or absolute position information of the individual patches. The entire sequence I^f is then fed into a transformer encoder consisting of multiple layers. Each layer of the encoder contains a self-attention module and a feed-forward network that allows the model to capture the information between image patches. After being processed by the image encoder, the output sequence $I^{f'}$ contains rich contextual information that can be further used for subsequent global alignment tasks and disease feature extraction processes. The whole process of image feature extraction can be formalized as follows:

$$X_I = \text{LN}(\text{SelfAttn}(X_{I-1}) + X_{I-1}) \quad (1)$$

$$X_I = \text{LN}(\text{FFN}(X_I) + X_I) \quad (2)$$

where SelfAttn stands for self-attention, LN stands for layer normalization, and FFN stands for feed-forward network [18].

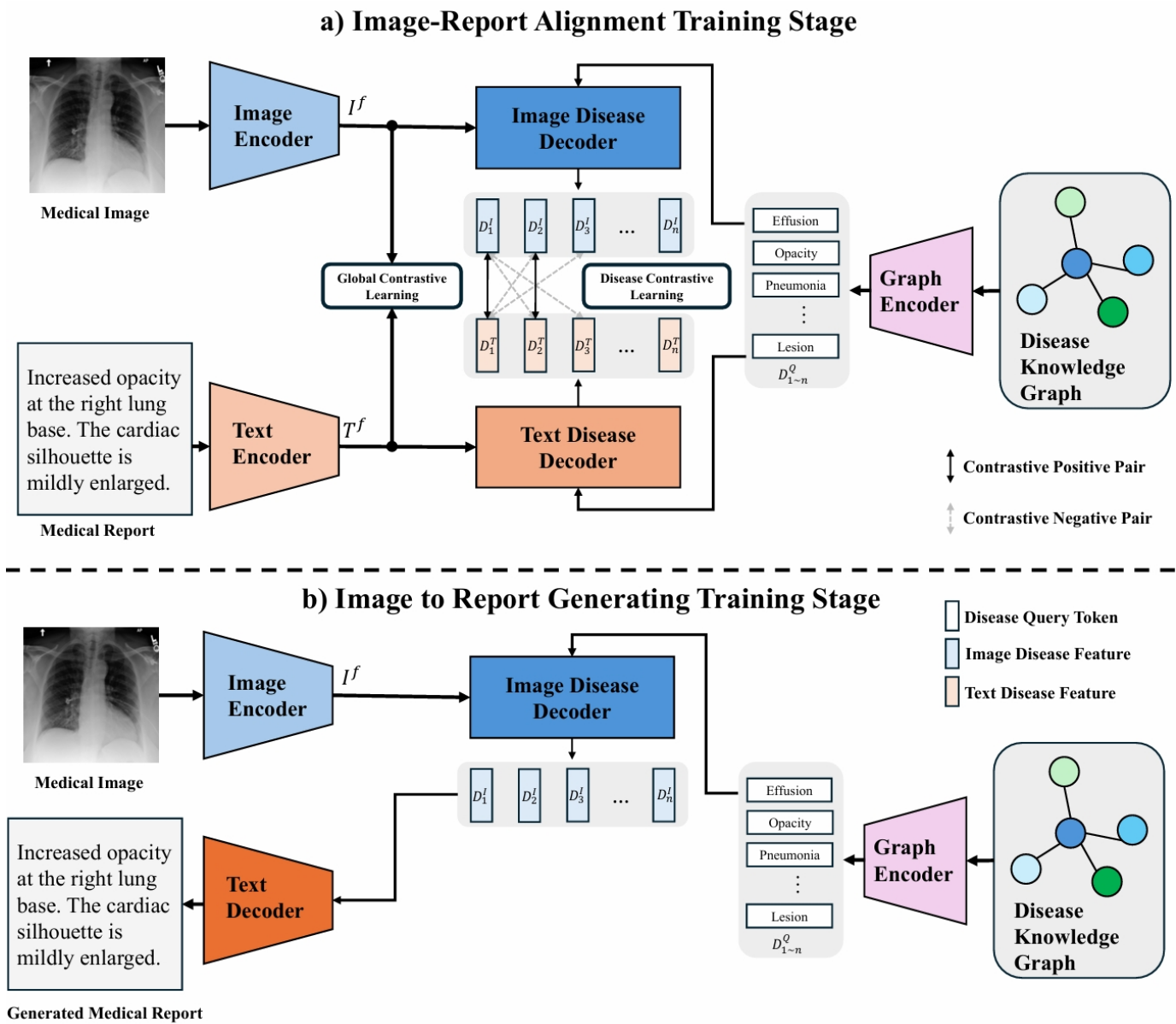


Figure 1. Illustration of our proposed DKA-RG framework.

Text Feature Extraction. We use BERT [16] as the text encoder as shown in Figure 2. Given a radiology report text T , it is first cleaned to remove non-standard characters to fit the input requirements of the model. Then, the text is segmented into multiple tokens, and each token t_i is mapped to a high-dimensional feature vector T_i^f through an embedding layer. Similar to ViT [15], each text token is also summed with a position embedding to insert position information. The entire sequence T^f is then fed into a Transformer encoder consisting of multiple layers. After processing by the Transformer encoder, the output sequence $T^{f'}$ is used for subsequent global alignment tasks and disease feature extraction processes.

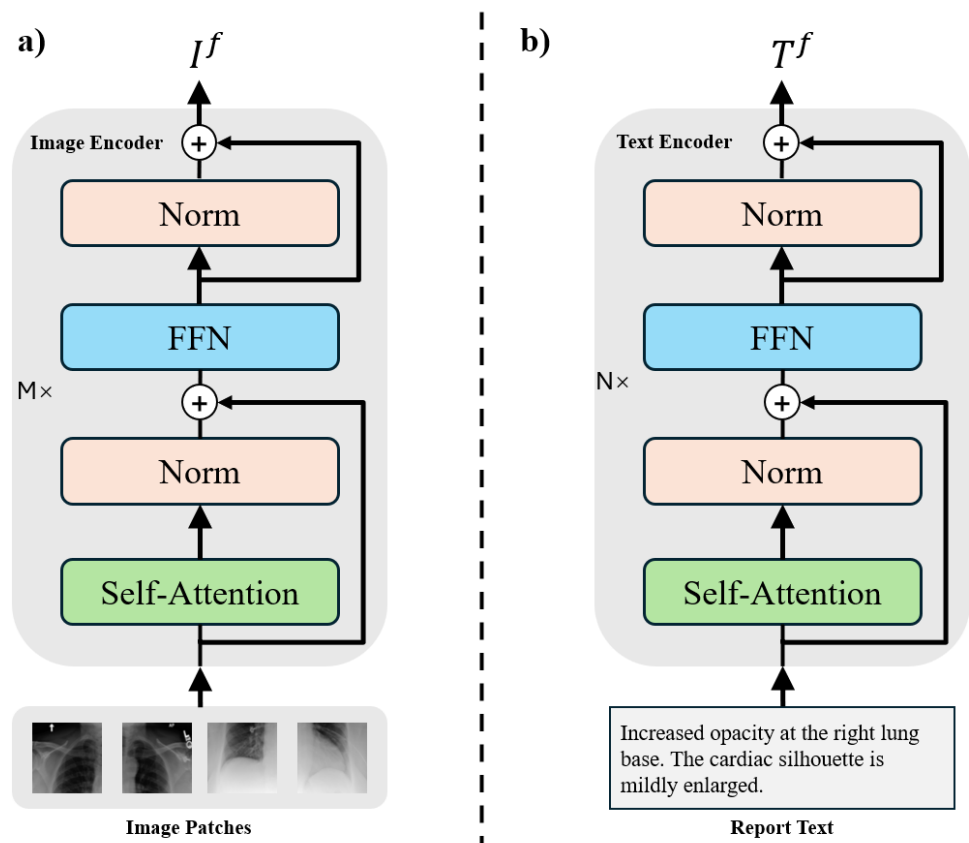


Figure 2. Architecture for image and text encoders. (a) shows the image encoder, which is based on the Vision transformer structure. It has M layers (12 in our specific implementation) of transformer blocks. (b) shows the text encoder, which is based on the BERT structure and consists of N layers (also implemented as 12) of transformer blocks.

3.2. Disease Knowledge Graph Embedding

Disease Knowledge Graph. The medical domain is different from the general-purpose domain in that it is filled with a large amount of specialized knowledge. In order to introduce medical expertise into the modeling framework, we follow some existing work to construct a disease knowledge graph, i.e., a graph structure of the 20 diseases that appear with high frequency in chest radiograph reports according to the distribution of their anatomical locations, as shown in Figure 3. For example, the heart may have cardiac enlargement, and the lungs may have edema, emphysema, etc. Knowledge mapping can reveal potential associations between different diseases, such as the co-occurrence, causation, and similarity of pathomechanisms, which can help to understand the complexity and diversity of the diseases and to identify co-morbid conditions.

Graph node initialization. First, the names of 20 diseases are encoded using the BERT [16] model to obtain an initial disease node representation:

$$\mathbf{h}_i = \text{BERT}(D_i), \quad i = 1, 2, \dots, N_{\text{disease}} \quad (3)$$

where D_i represents the name of the i th disease and N_{disease} is the total number of diseases (20 in our implementation).

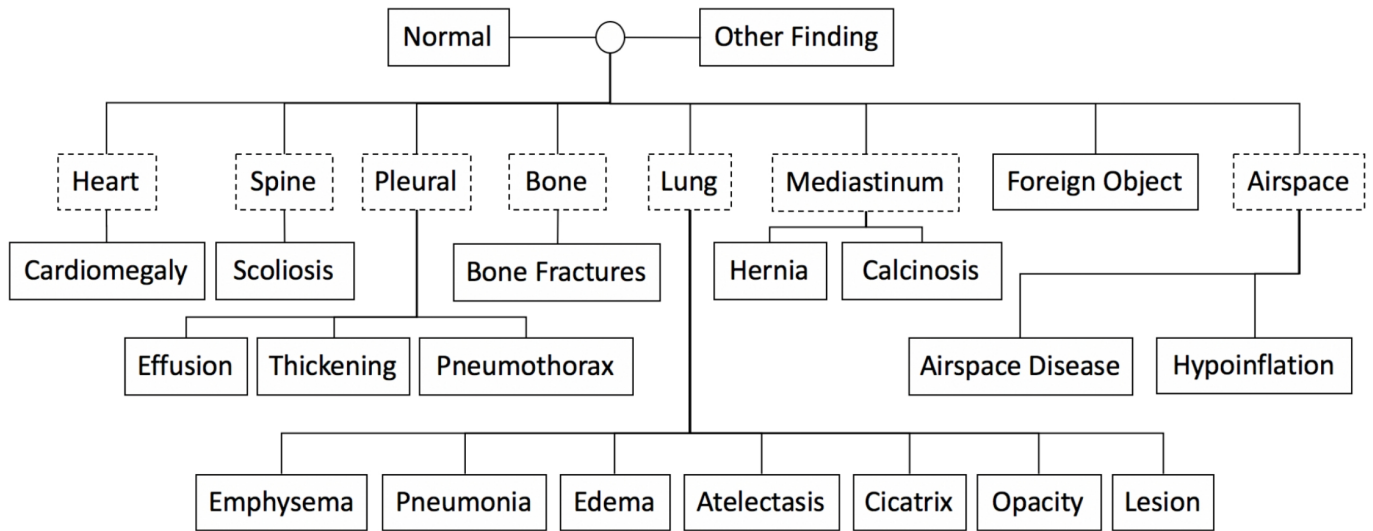


Figure 3. Knowledge graph containing 20 chest radiograph diseases and their relationships (solid borders are diseases, dashed borders are anatomical locations).

Neighborhood information aggregation based on graph attention networks (GAT).

The disease nodes are considered as nodes in the graph, and the graph embedding of these nodes is performed using GAT, as shown in Figure 4. Let $\mathcal{N}(i)$ be the set of neighboring nodes of node i . In GAT, the new embedding representation $\mathbf{h}'_i$ of node i is computed by aggregating the representations of neighboring nodes:

$$\mathbf{h}'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W} \mathbf{h}_j \right) \quad (4)$$

where σ is a nonlinear activation function, $\mathbf{W}$ is a learnable weight, and α_{ij} is an attention weight indicating the importance of neighboring node j to node i . The attention weight α_{ij} is computed from the representation of node i and node j :

$$e_{ij} = \text{LeakyReLU} \left(\mathbf{a}^T [\mathbf{W} \mathbf{h}_i \parallel \mathbf{W} \mathbf{h}_j] \right) \quad (5)$$

where $\mathbf{a}$ is a learnable weight vector, and $\parallel$ denotes the vector's join operation. Next, the attention weights of all neighboring nodes are normalized:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \quad (6)$$

Through the above steps, the new embedding representation $\mathbf{h}'_i$ of each node can be calculated. Finally, the final disease feature vector is obtained by propagating through a multi-layer GAT network. Let the GAT network have L layers (5 layers in our implementation; the representation of node i in the l th layer is

$$\mathbf{h}_i^{(l)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)} \right) \quad (7)$$

After the propagation of L layers, the final disease feature vectors $\mathbf{h}_i^{(L)}$ are obtained. These feature vectors will be used as disease query tokens to extract disease information from images and reports.

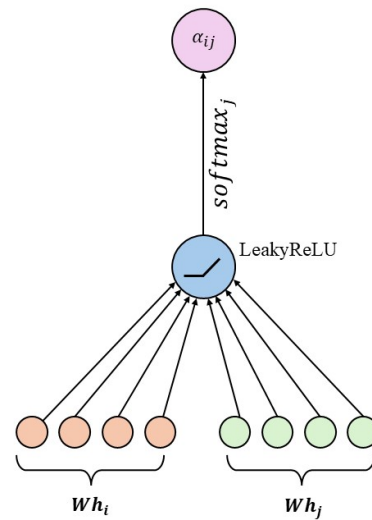


Figure 4. Aggregate node information using graph attention mechanism. The node i and node j are transformed and computed with learnable weights $\mathbf{W}$, respectively, and after activation function LeakyReLU activation and softmax function, the attention weights α_{ij} are obtained.

3.3. Image–Report Alignment

Global Contrastive Learning. Contrastive learning can effectively align data (e.g., images and text) from different modalities, making information with the same semantics closer in the representation space of different modalities [21,22,36–38]. Through contrastive learning, images and text can be mapped to a unified representation space, thus achieving semantic consistency across modalities. Contrastive learning achieves efficient representation learning by bringing feature distances between pairs of positive samples (images and corresponding texts) closer together and pushing feature distances between pairs of negative samples (images and mismatched texts) farther apart. This representation not only performs well in cross-modal tasks but can also be used in downstream tasks to improve the generalization ability of the model.

As shown in Figure 5a,b, after global alignment training, the features obtained from the encoder of the image and report will move from a chaotic state to an aligned state, which makes the extracted features of the image richer in valuable information.

We follow [22] using the momentum contrastive learning approach. Given an image I and a corresponding report T , we extract the features mentioned above to obtain a sequence of image and text features $I^{f'}$ and $T^{f'}$ to be averaged and mapped to a uniform dimension to obtain the global features I_g^f and T_g^f . Based on the global features, we can compute the cosine similarity $s(I, R)$ between image and text. Then, we use InfoNCE loss [22] to compute image-to-text similarity and text-to-image similarity:

$$p_k^{\text{I2T}}(I) = \frac{\exp(s(I, T_k)/\tau)}{\sum_{k=1}^K \exp(s(I, T_k)/\tau)} \quad (8)$$

$$p_k^{\text{R2I}}(T) = \frac{\exp(s(T, I_k)/\tau)}{\sum_{k=1}^K \exp(s(T, I_k)/\tau)} \quad (9)$$

where τ is a learnable parameter, and K represents the length of the queue in momentum-based contrastive learning. The loss function for global alignment is defined as follows:

$$\mathcal{L}_{\text{global}} = \frac{1}{2} \mathbb{E}_{(I,T) \sim D} [\mathcal{L}_{\text{CE}}(\mathbf{y}^{\text{I2T}}(I), \mathbf{p}^{\text{I2T}}(I)) + \mathcal{L}_{\text{CE}}(\mathbf{y}^{\text{T2I}}(T), \mathbf{p}^{\text{T2I}}(T))] \quad (10)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss function, and $\mathbf{y}^*(\cdot)$ is the ground truth of similarity based on constructed positive–negative pairs.

After global contrastive learning, the features of the two modalities, image and text, reach a coarse-grained alignment. This alignment facilitates the ability of the image encoder to extract medical image features.

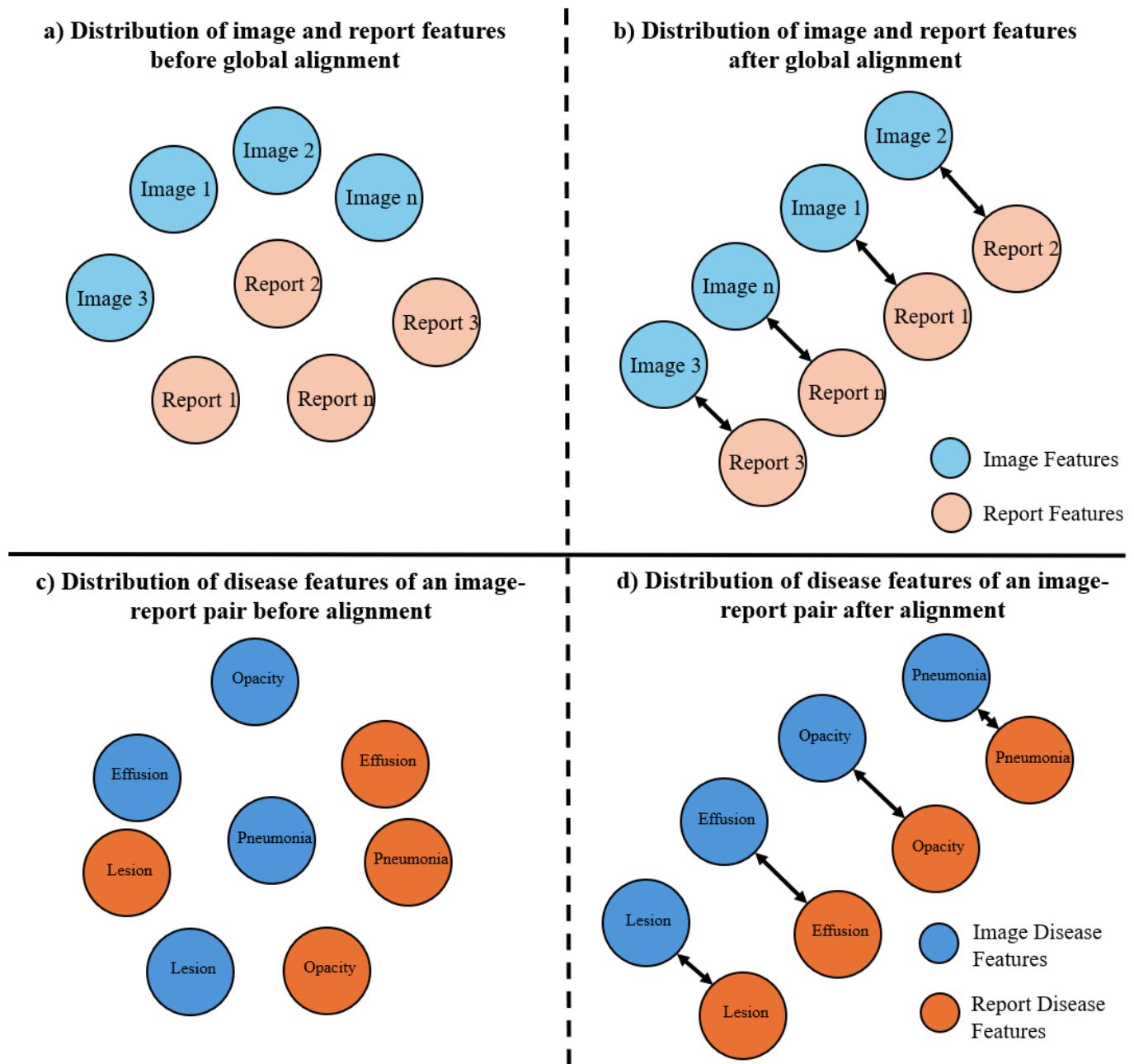


Figure 5. Demonstration of global alignment and disease alignment. The four subfigures of (a–d) show the feature distribution states before and after alignment. The feature distribution before alignment is chaotic, while the matched features after alignment are pulled closer, which promotes the learning ability of the model.

Disease Contrastive Learning. Although global contrastive learning has significant advantages in aligning medical images and report texts to promote semantic consistency across modalities, its alignment is still coarse-grained. Global contrastive learning focuses on the overall semantic alignment of images and text, often ignoring the semantic alignment of specific disease details. This approach is difficult to capture the relationship between the tiny features of lesions in images and the fine descriptions in the report text, thus

limiting the accurate analysis and diagnosis of subtle lesions in complex medical scenarios. Fine-grained alignment, which can more effectively link specific lesion regions in images and corresponding text descriptions, is the key to further improve the alignment of medical images and report texts.

As shown in Figure 5, through contrastive learning, the same disease features are brought closer together, thereby achieving semantic alignment at the disease level and improving the model's ability to extract disease features.

To address the limitations of global alignment, we propose a knowledge-based disease contrastive learning. Specifically, we construct an image disease decoder and a text disease decoder based on the Transformer decoder structure for extracting disease-specific features from images/text, as shown in Figure 6. In order to extract the disease information in images and texts, we use the disease embedding vectors obtained in the previous section as disease query tokens and interact with the image or text feature tokens through the cross attention module, so that the disease information in a specific image or text can be extracted eventually. For an image and report text pair, they essentially contain the same disease semantic information and thus should be queried for similar information through the disease decoder, so we construct disease-level contrastive learning. Specifically, we obtain the image/text disease features $\mathbf{D}_{1 \sim n}^I$ and $\mathbf{D}_{1 \sim n}^T$ through the disease decoder. Disease contrastive loss is constructed in the following form:

$$\mathcal{L}_{disease} = - \sum_{i=1}^n \log \frac{\exp(\text{sim}(\mathbf{D}_i^I, \mathbf{D}_i^T))}{\exp(\text{sim}(\mathbf{D}_i^I, \mathbf{D}_i^T)) + \sum_{j \neq i} \exp(\text{sim}(\mathbf{D}_i^I, \mathbf{D}_j^T))} \quad (11)$$

where sim is the cosine similarity function, and n is the total number of diseases.

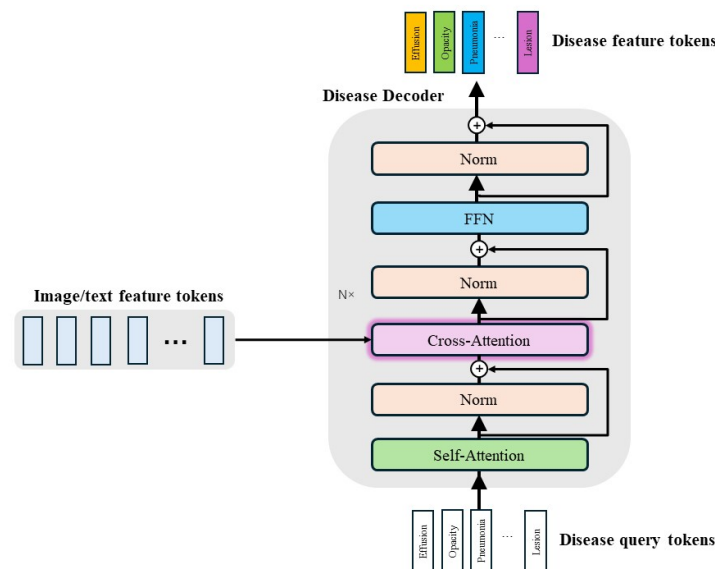


Figure 6. Disease decoder based on Transformer decoder structure. Disease query tokens are taken as input and made to interact with image or text features to extract the relevant disease features in the image/text.

Our total image text alignment loss is the sum of the global contrastive learning and disease contrastive learning losses:

$$\mathcal{L}_{alignment} = \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{disease} \quad (12)$$

where λ_1 and λ_2 are the weighting factors that control the two losses (set to 1 by default in the experiment).

3.4. Image to Report Generation

After the alignment of image and text, the image encoder and image disease decoder can fully extract disease features from medical images. So we build a text decoder to transform the disease information extracted from the images into a final report. Specifically, we use a text decoder based on a transformer decoder. After splicing a special start token $\langle \text{bos} \rangle$ in front of the report text, it is fed into the text decoder. At the same time, we use the disease features $\mathbf{D}_{1 \sim n}^I$ extracted by the image disease decoder as the input to the cross attention module of the text decoder, as shown in Figure 1b. Finally, we perform text generation autoregressive loss computation:

$$\mathcal{L}_{\text{generation}} = - \sum_{i=1}^N \log P(x_i | x_{<i}, \mathbf{D}^I) \quad (13)$$

where $\mathbf{D}^I$ represents the image disease features, and N is the number of report tokens.

4. Experiments

4.1. Datasets

There Author : *Confirmed. No software mentioned in the paper.* are two widely used chest radiograph datasets in computerized medicine: IU-Xray [20] and MIMIC-CXR [19]. We performed extensive experiments on these two datasets.

IU-Xray. The IU-Xray dataset [20] is widely used in the field of radiology report generation. The amount of data contained in the IU-Xray is shown in Table 2. Each report in IU-Xray either corresponds to a frontal view or a combination of a frontal and a lateral view. Specifically, we divide the dataset into training/validation/testing sets in the ratio 7:1:2 based on the work [39].

MIMIC-CXR. The MIMIC-CXR dataset [19] is currently the largest chest X-ray report dataset. The specific number of samples in the MIMIC-CXR dataset is shown in Table 2. We directly use the official way of dividing the dataset, in line with most methods. The dataset can be used to support a wide range of medical tasks, including medical image understanding and report generation. Due to its large data volume, this dataset enables our method to mine richer disease information.

Table 2. Statistics on the number of training, validation, and test sets for both datasets and the average length of the reports.

Datasets	IU-Xray			MIMIC-CXR		
	TRAIN	VALIDATION	TEST	TRAIN	VALIDATION	TEST
Num of Images	5226	748	1496	368,960	2991	5159
Num of Reports	2770	395	790	222,758	1808	3269
Num of Patients	2770	395	790	64,586	500	293
AVG. LEN.	37.56	36.78	33.62	53.00	53.05	66.40

4.2. Evaluation Metrics

We use the most common natural language generation (NLG) metrics and clinical efficacy (CE) metrics to assess our model's performance.

BLEU. The BLEU [40] metric is an indicator originally used to evaluate the quality of machine translation. It mainly calculates the similarity between sentences from the n-gram similarity between the generated text and the reference text. BLEU mainly calculates the quality of the generated text from the perspective of precision, so the BLEU metric is also widely used to evaluate the precision of radiology report generation. A high BLEU metric represents a high percentage of generated reports that are consistent with ground-truth reports.

ROUGE-L. Unlike the BLEU metric, the ROUGE metric [41] measures the proximity of the machine-generated text to the reference text in terms of recall. In addition, ROUGE-L differs from BLEU in calculating n-gram similarity, which calculates the longest common

subsequence of the ground-truth report and the generated report. Therefore, the ROUGE-L metric can be used to measure whether our generated report successfully recalls the content in the ground-truth report. A good generated report should maintain a high level of both BLEU metrics and ROUGE-L metrics.

METEOR. METEOR [42] is another metric for evaluating text generation. It calculates the similarity between the generated text and the reference text in terms of word-level accuracy and recall, as well as penalties for order. In addition, cases such as synonyms are not considered in the BLEU metric and ROUGE, whereas the METEOR metric takes into account the exact matching of words and also the matching of stems, synonyms, and other linguistic variants.

CIDEr. The CIDEr metric [43] is specifically designed to evaluate the quality of image descriptions. Compared to the evaluation metrics BLEU and ROUGE, which are commonly used for text translation, CIDEr is closer to the principle of the human ability to judge whether two sentences are similar, because it takes into account the frequency of occurrence of n-grams in a sentence, thus reducing some parts that are high-frequency but contain little information. Thus, CIDEr focuses more on evaluating that the model has extracted more unique information from the image rather than some generic utterances.

Clinical Efficacy. Clinical efficacy metrics have recently been introduced to assess the accuracy of predictive radiology reports. The CheXbert labeling tool described in [44] is used to annotate prediction reports and ground-truth reports for 14 different medical diseases. Subsequently, classification metrics such as F1-score, precision, and recall were computed to assess the validity of the generated reports in describing the anomalies. Notably, since the IU-Xray provider does not employ CheXbert for label generation, CE metrics are only reported for the MIMIC-CXR dataset.

4.3. Implementation Details

For the image encoder, we use the pre-trained 12-layer ViT/B16 [15] at ImageNet dataset [45]. For the text encoder, we use the pre-trained 12-layer BERT [16]. For the image/text disease decoder, we use a 3-layer transformer decoder [18] structure with random parameter initialization. For the text decoder, we use a 6-layer transformer decoder structure and initialize it with BERT weights. There are two stages in the training process of our method. Stage 1 performs image–text alignment training with 20/5 epochs on IU-Xray and MIMIC-CXR datasets, respectively. Stage 2 performs image-to-report generation training, which retains the image encoder, image disease decoder, and graph encoder that have been trained in stage 1 and introduces the text decoder for 10/3 epochs of training on the two datasets, respectively. We use a NVIDIA 4090 GPU (produced by NVIDIA Corporation, Santa Clara, CA, USA) to train our model. We set the batch size to 16 and the learning rate to 1×10^{-4} with cosine annealing decay strategy. The optimizer uses RAdam [46] and sets the weight decay to 0.95.

4.4. Main Results

NLG Results. Table 3 shows the results of our method on the two used datasets in comparison with other methods. On the IU X-Ray dataset, the DKA-RG method is the best on most of the metrics, especially on the BLEU and CIDEr metrics, where the improvement is more obvious. In addition, the ROUGE-L metric is slightly lower than the DCL method but still reaches 0.376, showing a strong text generation capability. The METEOR score of 0.201 is the highest among all the methods, and the CIDEr score is even as high as 0.621, which demonstrates that our method generates reports that are more relevant to the images. The DKA-RG method performance is also stable on the MIMIC-CXR dataset. Its scores on BLEU-1~4 are 0.395, 0.254, 0.165, and 0.122, respectively, which are all ahead of other methods. Although the ROUGE-L metric is slightly lower than the 0.285 of the KiUT method, it also reaches 0.282, showing better text relevance. Meanwhile, the CIDEr score is 0.303, which again verifies its superiority in generating high-quality reports.

Table 3. Comparison of our method with other methods on multiple language generation metrics. The best results are in **boldface**. B-n is the abbreviation of BLEU-n.

Dataset	Methods	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
IU X-ray	R2Gen [39]	0.470	0.304	0.219	0.165	0.371	0.187	-
	CA [34]	0.492	0.314	0.222	0.169	0.381	0.193	-
	PPKED [47]	0.483	0.315	0.224	0.168	0.376	-	0.351
	R2GenCMN [48]	0.475	0.309	0.222	0.170	0.375	0.191	-
	KnowMat [24]	0.496	0.327	0.238	0.178	0.381	-	0.382
	MSAT [49]	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	METransformer [33]	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	DCL [12]	-	-	-	0.163	0.383	0.193	0.586
	DKA-RG (Ours)	0.496	0.328	0.239	0.182	0.376	0.201	0.621
MIMIC -CXR	R2Gen [39]	0.353	0.218	0.145	0.103	0.277	0.142	-
	CA [34]	0.350	0.219	0.152	0.109	0.283	0.151	-
	PPKED [47]	0.360	0.224	0.149	0.106	0.284	0.149	-
	R2GenCMN [48]	0.353	0.218	0.148	0.106	0.278	0.142	-
	KnowMat [24]	0.363	0.228	0.156	0.115	0.284	-	0.203
	MSAT [49]	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	KiUT [35]	0.393	0.243	0.159	0.113	0.285	0.160	-
	DCL [12]	-	-	-	0.109	0.284	0.150	0.281
	DKA-RG (Ours)	0.395	0.254	0.165	0.122	0.282	0.154	0.303

Clinical Efficacy. Table 4 shows the comparison results of our method with other methods in clinical efficacy indicators, which corresponds to the overall classification performance of the model for diseases. Compared with methods such as R2Gen [39], R2GenCMN [48], KnowMat [24], KiUT [35], and DCL [12], our DKA-RG shows a higher level of performance, with precision, recall, and F1-score of 0.489, 0.384, and 0.430, respectively, far ahead of other methods. This shows that our DKA-RG has a good ability to detect diseases in images and reflect them in reports. In addition, we also explore the F1-score of DKA-RG on specific diseases, as shown in Figure 7. As can be seen from the figure, our method achieves a higher F1-score than the baseline method (without global-level and disease-level alignment) for most diseases. Our method has achieved significant improvements in cardiomegaly, edema, consolidation, pneumonia, and pneumothorax. This improvement shows that our designed global-level and disease-level alignment and knowledge graph enhancement can enable the model to have better disease mining capabilities.

Table 4. Comparative results between our DKA-RG method and other methods on clinical efficacy metrics. Since the IU-Xray dataset [20] does not provide its ground truth on clinical efficacy, we only perform this comparison experiment on the MIMIC-CXR dataset [19]. The best results are in **boldface**.

Methods	P	R	F1
R2Gen	0.333	0.273	0.276
R2GenCMN	0.334	0.275	0.278
KnowMat	0.458	0.348	0.371
KiUT	0.371	0.318	0.321
DCL	0.471	0.352	0.373
DKA-RG (ours)	0.489	0.384	0.430

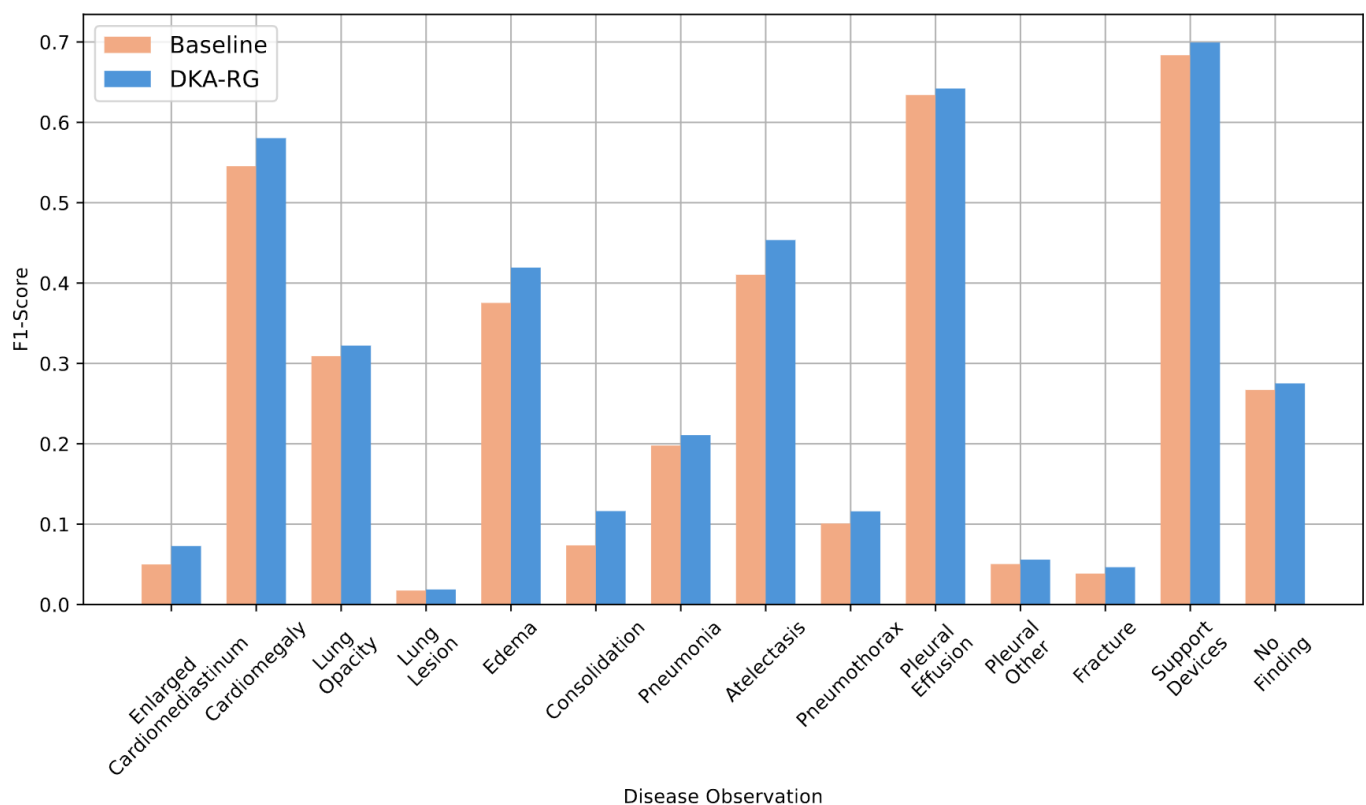


Figure 7. Clinical efficacy F1-score on different diseases for the baseline and our DKA-RG method.

4.5. Ablation Study

Table 5 shows the performance of our model in different module combinations. Specifically, we set up 5 different situations. The base represents the training of image-to-report generation directly using only one image encoder and text decoder without any alignment. Setting (a) performs global contrastive learning on the basis of base setting, setting (b) introduces disease contrastive learning on the basis of base setting, setting (c) further introduces knowledge graph on the basis of setting (b), and setting (d) is our complete DKA-RG method.

Table 5. Ablation results on IU-Xray dataset. Base stands for a method that uses only one image encoder and one text decoder. Global CL stands for global contrastive learning. Disease CL stands for disease contrastive learning. The check mark indicates that the corresponding module is used.

Settings	Global CL	Disease CL	Knowledge Graph	BLEU-4	ROUGE-L	METEOR	CIDEr	AVG.Δ
Base				0.166	0.365	0.194	0.454	-
(a)	✓			0.178	0.373	0.201	0.547	+10.2%
(b)		✓		0.175	0.368	0.197	0.541	+8.7%
(c)		✓	✓	0.181	0.373	0.203	0.589	+14.2%
(d)	✓	✓	✓	0.182	0.376	0.201	0.621	+17.0%

Effect of global contrastive learning (Global CL). We only use global contrastive learning in setting (a). Global alignment already improves performance more than the baseline method (without any alignment). Specifically, BLEU-4 increases from 0.166 to 0.178, ROUGE-L from 0.365 to 0.373, METEOR from 0.194 to 0.201, and CIDEr from 0.454 to 0.547. This results in an average improvement (AVG.Δ) of +10.2%. This demonstrates that global alignment can help the model establish semantic associations between medical images and reports. Since the final report needs to embody the same information as the image, this

ability to globally semantically correlate improves report generation and produces reports that are better aligned to the global features of the image.

Effect of disease contrastive learning (Disease CL). In setting (b), we introduce disease contrastive learning, which is slightly lower than the performance improvement from global contrastive learning in (a), but still considerably higher than the metrics of the baseline method. Specifically, BLEU-4 improves to 0.175, ROUGE-L to 0.368, METEOR to 0.197, and CIDEr to 0.541, with an AVG. Δ of +8.7%. This illustrates the ability of disease contrastive learning to help the model align the disease features in the images and reports so that this disease information is reflected in the final report generation session.

Effect of disease knowledge graph. In setting (c), we combine disease contrastive learning and the knowledge graph to enable the model to refer to relevant medical disease knowledge when performing disease alignment. This configuration shows significant improvements with BLEU-4 at 0.181, ROUGE-L at 0.373, METEOR at 0.203, and CIDEr at 0.589, resulting in an AVG. Δ of +14.2%. The interconnections of multiple diseases are represented in the medical knowledge graph, which introduces a priori knowledge for the model to perform disease alignment, thus improving the training stability of the model and avoiding the learning of useless noise.

Effect of combination of three modules. In setting (d), which is our final method, we introduce global contrastive learning, disease contrastive learning, and the knowledge graph at the same time and achieve a significant improvement in average performance (+17.0%). This shows that global-level alignment and knowledge-enhanced disease-level alignment can promote each other, and it also shows that alignment from different angles can improve the final model effect.

4.6. Qualitative Experiment

In order to verify the effectiveness of our method in real scenarios, we show two reports generated by our method and the baseline method, respectively, and the corresponding ground-truth reports in Figure 8.

In the first example, the ground-truth report states “cardiomegaly without acute cardiopulmonary process”, which is an exact match to the report generated by our method. Although the baseline method states “no acute cardiopulmonary process”, it omits “cardiomegaly”, which leads to inaccurate results. In addition, in the finding section of the report, our DKA-RG also describes that “cardiac silhouette is moderately enlarged”, which is omitted by the baseline method. This illustrates the ability of our method to generate more detailed reports when diagnosing images.

In the second example, both the baseline method and our method correctly point out “mild interstitial pulmonary edema” and “pleural effusion” in the impression section of the report. In the finding section, both the baseline method and our method notice the symptom of “mild cardiomegaly”. However, the report generated by the baseline method is too brief and misses some information. The ground-truth report mentions “tiny right pleural effusion”, which corresponds to the “small bilateral pleural effusion” mentioned by our method. This example once again demonstrates that our DKA-RG method can generate reports that are highly consistent with the ground-truth reports.

In both examples, the DKA-RG method can produce more detailed and accurate radiology reports compared to the baseline method. In generating reports, our method always captures the details of the disease in the medical images. This shows that the disease-level contrastive learning we designed has a good effect. These examples show the potential of our DKA-RG method in generating high-quality radiology reports, which is of great significance to promoting diagnostic decisions in the medical field.

Figure 9 shows more examples of comparing the medical reports generated by our method with ground truth. In most cases, our method can generate report content that is consistent with the ground truth, but there are also a few examples of errors. As shown in the 7th and 8th examples in Figure 9, this also indicates that the current medical report

automatic generation system has not yet reached a very perfect level, but it can still be used as an auxiliary tool for human physicians.

Ground Truth



impression : **cardiomegaly without acute cardiopulmonary process** . findings : **the lungs are clear of consolidation effusion or pneumothorax** . left chest wall dual lead pacing device is again seen . **moderate cardiomegaly is again noted** . upper thoracic dextroscoliosis is seen . no acute fracture identified based on this nondedicated exam . surgical clips seen in the upper abdomen .

Baseline

impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest . **the lungs are clear of focal consolidation effusion or pulmonary vascular congestion** . no acute osseous abnormality is identified . hypertrophic changes are seen in the spine without acute osseous abnormality .

Ours

impression : **cardiomegaly without acute cardiopulmonary process** . findings : **the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen** . **the cardiac silhouette is moderately enlarged** . the aorta is slightly tortuous . dual lead left - sided pacemaker is seen with leads extending to the expected positions of the right atrium and right ventricle .

Ground Truth



impression : **new mild interstitial edema and tiny right pleural effusion** . findings : **interstitial prominence** has increased compared to prior suggestive of mild edema . **no focal consolidation or pneumothorax is detected** . **tiny right pleural effusion** appears new compared to prior . **heart and mediastinal contours appear stable with mild cardiomegaly** .

Baseline

impression : **mild interstitial pulmonary edema and small bilateral pleural effusions** . findings : **moderate enlargement of the cardiac silhouette** . the aorta is tortuous and calcified . **no pneumothorax is seen** .

Ours

impression : **mild interstitial pulmonary edema and small bilateral pleural effusions** . findings : **moderate enlargement of the cardiac silhouette is re - demonstrated** . the aorta is tortuous and diffusely calcified . **there is mild interstitial pulmonary edema** new in the interval with **small bilateral pleural effusions** . **no pneumothorax is seen** . there are no acute osseous abnormalities .

Figure 8. Examples of our DKA-RG and baseline generated radiology reports. Sentences of the same color indicate that they point to the description of the same disease symptom.

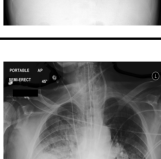
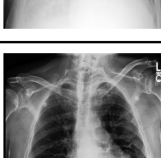
	Image	Ground Truth	Ours
1		impression : small left pleural effusion and improving atelectasis but no pneumothorax . findings : the cardiomediastinal and hilar contours are normal . subtle linear horizontally oriented opacities in the left costophrenic angle appear improved compared to prior exams and likely reflect the sequelae of resolving atelectasis .	impression : small left pleural effusion . findings : the cardiomediastinal and hilar contours are normal . there is no pneumothorax . there is a small left pleural effusion . there is no focal consolidation concerning for pneumonia . pulmonary vasculature is within normal limits . there is no pneumothorax .
2		impression : no acute cardiopulmonary process . findings : the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable . the patient is status post median sternotomy with the superior most 2 sternotomy wires again seen to be fractured .	impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest were obtained . the patient is status post median sternotomy and cabg . no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are stable and unremarkable . no overt pulmonary edema is seen .
3		impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest . the lungs are clear of consolidation or effusion . right pleural based thickening at the base laterally is again seen . there is no evidence of pulmonary vascular congestion . cardiomediastinal silhouette is stable in . no acute osseous abnormality detected .	impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest were obtained . the patient is status post median sternotomy and cabg . the cardiac and mediastinal silhouettes are stable with the cardiac silhouette enlarged . no focal consolidation pleural effusion or evidence of pneumothorax is seen .
4		low lung volumes are present . this accentuates the size of the cardiac silhouette which is likely mildly enlarged . mediastinal and hilar contours are likely within normal limits . a right brachiocephalic venous stent is re - demonstrated . there is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion .	lung volumes are low . this accentuates the size of the cardiac silhouette which appears mildly enlarged . the mediastinal and hilar contours are unremarkable . there is crowding of the bronchovascular structures but no overt pulmonary edema is present . patchy opacities in the lung bases likely reflect atelectasis .
5		impression : no free air under the diaphragms . low lung volumes with bibasilar atelectasis . probable small left pleural effusion . findings : there are low lung volumes . the heart size is normal . the aorta remains tortuous . there is crowding of the bronchovascular structures but no overt pulmonary edema . linear opacities at the lung bases likely reflect atelectasis .	impression : low lung volumes with bibasilar atelectasis . findings : lung volumes are low . the aorta is tortuous and diffusely calcified . mediastinal and hilar contours are otherwise unremarkable . there is crowding of the bronchovascular structures but no overt pulmonary edema is present .
6		impression : no significant interval change . findings : frontal and lateral views of the chest were obtained . there are relatively low lung volumes . mild elevation of the right hemidiaphragm persists . there is persistent right base atelectasis . no new focal consolidation is seen . there is no pleural effusion or pneumothorax .	impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest were obtained . there are relatively low lung volumes . given this no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable . no overt pulmonary edema is seen . no displaced fracture is seen .
7		impression : lines and tubes in place . increased pulmonary edema with right upper lobe and bibasilar consolidations . findings : the endotracheal tube tip sits 5 cm above the carina . a left - sided ij central venous catheter tip sits in the left brachiocephalic vein . the right - sided ij central venous catheter tip sits in the upper svc . the heart size is large but stable .	the monitoring and support devices are constant . constant appearance of the lung parenchyma and of the cardiac silhouette . no new parenchymal opacities . no larger pleural effusions . no pneumothorax . no overt pulmonary edema . unchanged appearance of the hilar and mediastinal contours .
8		heart size remains mildly enlarged . aortic knob is calcified . mediastinal and hilar contours are unchanged . previously noted left upper lobe mass appears more vague with surrounding ill - defined opacity possibly related to infection .	heart size is normal . mediastinal and hilar contours are unremarkable . pulmonary vasculature is normal . linear opacities within the left lung base likely reflect subsegmental atelectasis . no focal consolidation pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .

Figure 9. More comparisons between the reports generated by DKA-RG and the ground truth, where bold text indicates that the generated reports contain the same content as the ground truth. Our method still has the possibility of errors, as indicated by the highlighted parts in the image that describe symptoms that are inconsistent with the ground truth.

5. Discussion

5.1. Main Novelty

As mentioned above, the core innovation of our work is the alignment of features of medical images and report texts with the help of the medical knowledge graph. Previous work often only performs coarse-grained alignment at the global level or directly trains end-to-end image-to-report text generation, which is suitable for general-purpose domains but not for domains such as medical diagnosis that require more detailed processing of images and text. Our approach takes into account the specificity of medical diagnosis and introduces the medical knowledge graph as an additional knowledge supplement to help the model mine medical features. At the same time, we additionally design image-report disease-level feature alignment for disease disorders that has not been considered in previous approaches, which facilitates the feature extraction capability of medical images and reports at a finer-grained level, thus improving the accuracy of the final medical image-to-report generation.

5.2. Limitations

Our method also has some limitations. First, although our method is theoretically generalizable, we have only experimented on the chest image dataset because open datasets in the medical domain are very scarce. Other types of medical images tend to be smaller in size and lack high-quality physician-written reports, making it difficult to conduct experiments. Second, due to the limitation of computational resources, our method does not introduce larger scale models (e.g., large models with 7B or 13B parameters) to further enhance the modeling effect. Finally, the knowledge graph used in our method is fixed and does not cover all types of disease, which is a point that can be optimized in the future.

5.3. Practical Application

Our radiology report automatic generation system has high practicality. Although the automatic generation of radiology reports is still in the early stages of exploration, our method relies less on computing resources and can easily be inherited into existing healthcare systems. For example, for an existing chest radiograph scanning system, we can install our method based on its computer environment. Our method uses GPU during the training phase but can rely entirely on CPU during the usage phase. The chest X-ray scanning system takes the scanned images as input to our system, and ultimately, our system outputs a medical report in text format for future reference by professional doctors. Our method is compatible with most x86-based computers and mainstream operating systems.

6. Conclusions and Future Works

6.1. Conclusions

In this work, we build a novel framework for automatic radiology report generation, DKA-RG. We use multi-layer contrastive learning to enhance the model's ability to extract disease features from radiology images. At the same time, we also construct a medical knowledge graph to introduce medical knowledge into the alignment process, making it easier for the model to understand the relationships between diseases and the anatomical locations of diseases. Our DKA-RG can ultimately generate more accurate radiology reports due to its ability to extract more detailed disease semantic information. Our multiple experiments on two X-ray datasets have demonstrated the effectiveness of our method, thus validating the necessity of global and disease-level alignment and medical knowledge.

The radiology report automatic generation system we have constructed can be widely applied in practical medical scenarios. Our method can further promote the work efficiency of modern medical practitioners, reduce medical labor costs, improve the standardization and accuracy of medical diagnosis, and thus promote the development of interdisciplinary medical and computer science.

6.2. Future Works

In the future, we will further optimize our method. Currently, to improve the training and inference efficiency of the model and reduce the resource consumption of the automatic report generation system when used in actual environments, we select moderately sized models (12-layer BERT and 12-layer ViT). Considering that the medical field attaches more importance to the fluency and accuracy of reports, we plan to use the medical large language model as our text decoder in the future. Through the internal expertise of the pre-trained medical large language model, our method should be able to further improve performance. In addition, we are also considering expanding this method to more medical fields in the future. We will further design methods to improve the efficiency of data utilization to make up for the problem of data scarcity in other medical fields, thereby enhancing the universality of our method.

Author Contributions: Methodology, H.Y.; validation, H.Y.; formal analysis, H.Y.; investigation, H.Y.; data curation, H.Y.; writing—original draft preparation, H.Y.; writing—review and editing, H.Y., W.W. and Y.H.; visualization, H.Y.; supervision, W.W. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Sichuan Transportation Science and Technology Program (No. 2018-ZL-02).

Data Availability Statement: The two datasets used in this paper can be found at the following publicly available websites: <https://www.physionet.org/content/mimic-cxr/2.0.0>, <https://iuhealth.org/find-medical-services/x-rays>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bayareh-Mancilla, R.; Medina-Ramos, L.A.; Toriz-Vázquez, A.; Hernández-Rodríguez, Y.M.; Cigarroa-Mayorga, O.E. Automated computer-assisted medical decision-making system based on morphological shape and skin thickness analysis for asymmetry detection in mammographic images. *Diagnostics* **2023**, *13*, 3440. [CrossRef]
2. Cui, H.; Hu, L.; Chi, L. Advances in computer-aided medical image processing. *Appl. Sci.* **2023**, *13*, 7079. [CrossRef]
3. Bruno, M.A.; Walker, E.A.; Abujudeh, H.H. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *Radiographics* **2015**, *35*, 1668–1676. [CrossRef]
4. Jing, B.; Xie, P.; Xing, E. On the automatic generation of medical imaging reports. *arXiv* **2017**, arXiv:1711.08195.
5. Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; Xu, D. When radiology report generation meets knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12910–12917.
6. Akhter, Y.; Ranjan, R.; Singh, R.; Vatsa, M.; Chaudhury, S. On AI-assisted pneumoconiosis detection from chest x-rays. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 6353–6361.
7. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
8. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2048–2057.
9. Bai, S.; An, S. A survey on automatic image caption generation. *Neurocomputing* **2018**, *311*, 291–304. [CrossRef]
10. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.
11. Yan, A.; He, Z.; Lu, X.; Du, J.; Chang, E.; Gentili, A.; McAuley, J.; Hsu, C.N. Weakly supervised contrastive learning for chest X-ray report generation. *arXiv* **2021**, arXiv:2109.12242.
12. Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; Chang, X. Dynamic Graph Enhanced Contrastive Learning for Chest X-ray Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3334–3343.
13. You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; Wu, X. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Proceedings, Part III 24; Springer: Berlin/Heidelberg, Germany, 2021; pp. 72–82.
14. Tanida, T.; Müller, P.; Kaissis, G.; Rueckert, D. Interactive and Explainable Region-guided Radiology Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7433–7442.

15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2020.11929.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
17. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; MIT Press: Cambridge, MA, USA, 2017; Volume 30.
19. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
20. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [[CrossRef](#)] [[PubMed](#)]
21. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning, Virtual*, 13–18 July 2020; pp. 1597–1607.
22. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; pp. 9729–9738.
23. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference, Durham, NC, USA, 5–6 August 2022*; pp. 2–25.
24. Yang, S.; Wu, X.; Ge, S.; Zhou, S.K.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. [[CrossRef](#)]
25. Yuan, J.; Liao, H.; Luo, R.; Luo, J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part VI 22*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 721–729.
26. Yi, X.; Fu, Y.; Liu, R.; Zhang, H.; Hua, R. TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 2152–2162. [[CrossRef](#)]
27. Gu, T.; Liu, D.; Li, Z.; Cai, W. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–6 January 2024*; pp. 7995–8004.
28. Li, Y.; Liang, X.; Hu, Z.; Xing, E.P. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems, Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018*; Curran Associates Inc.: Red Hook, NY, USA, 2018; Volume 31.
29. Nakaura, T.; Yoshida, N.; Kobayashi, N.; Shiraishi, K.; Nagayama, Y.; Uetani, H.; Kidoh, M.; Hokamura, M.; Funama, Y.; Hirai, T. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports. *Jpn. J. Radiol.* **2024**, *42*, 190–200. [[CrossRef](#)]
30. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
31. Tang, Y.; Wang, D.; Zhang, L.; Yuan, Y. An efficient but effective writer: Diffusion-based semi-autoregressive transformer for automated radiology report generation. *Biomed. Signal Process. Control* **2024**, *88*, 105651. [[CrossRef](#)]
32. Jorg, T.; Halfmann, M.C.; Stoehr, F.; Arnhold, G.; Theobald, A.; Mildenerberger, P.; Müller, L. A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports. *Insights Into Imaging* **2024**, *15*, 80. [[CrossRef](#)] [[PubMed](#)]
33. Wang, Z.; Liu, L.; Wang, L.; Zhou, L. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology* **2023**, *1*, 100033. [[CrossRef](#)]
34. Ma, X.; Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zou, Y.; Zhang, P.; Sun, X. Contrastive attention for automatic chest X-ray report generation. *arXiv* **2021**, arXiv:2106.06965.
35. Huang, Z.; Zhang, X.; Zhang, S. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; pp. 19809–19818.
36. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning, Virtual*, 18–24 July 2021; pp. 8748–8763.
37. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems, Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online, 6–14 December 2021*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 9694–9705.
38. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022*; pp. 12888–12900.

39. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating radiology reports via memory-driven transformer. *arXiv* **2020**, arXiv:2010.16056.
40. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
41. Chin-Yew, L. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004.
42. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
43. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
44. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv* **2020**, arXiv:2004.09167.
45. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
46. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the variance of the adaptive learning rate and beyond. *arXiv* **2019**, arXiv:1908.03265.
47. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13753–13762.
48. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal memory networks for radiology report generation. *arXiv* **2022**, arXiv:2204.13258.
49. Wang, Z.; Tang, M.; Wang, L.; Li, X.; Zhou, L. A medical semantic-assisted transformer for radiographic report generation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 655–664.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.