Expert Systems With Applications 269 (2025) 126394

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Recalibrated cross-modal alignment network for radiology report generation with weakly supervised contrastive learning

Xiaodi Hou [a,1], Xiaobo Li [b,1], Zhi Liu [b], Shengtian Sang [c], Mingyu Lu [a], Yijia Zhang [b,*]

[a] School of Artificial Intelligence, Dalian Maritime University, Dalian, 116026, China
[b] School of Information Science and Technology, Dalian Maritime University, Dalian, 116026, China
[c] Department of Radiation Oncology, Stanford University, Stanford, CA, USA

ARTICLE INFO

Keywords:
Automatic radiology report generation
Visual and textual data bias
Cross-modal alignment
Weakly supervised contrastive learning

ABSTRACT

Automatic radiology report generation is rapidly becoming an essential method for medical diagnosis and precision medicine, which will help clinical doctors make more informed decisions and achieve better results. Most previous studies mainly employ an encoder–decoder architecture and tend to focus more on the text generation part, which ignore the following problems: (1) visual and textual data bias; (2) inadequate cross-modal interaction. In this paper, we propose a **R**ecalibrated **C**ross-modal **A**lignment **N**etwork for radiology report generation with weakly supervised contrastive learning (RCAN). Specifically, we design a recalibrated visual extractor to extract critical abnormal features. To achieve effective alignment between medical images and reports, we develop a cross-modal gated memory matrix. We effectively map cross-modal information by introducing gating units to memorize interactions between different modalities selectively. In addition, to encourage more accurate image anomaly detection and recognition, we propose a new weakly supervised contrastive learning approach. We conduct extensive experimental verification on two real-world datasets, IU X-ray and MIMIC-CXR, and the results show that our model can more accurately identify abnormal features in images and generate more continuous medical reports. Importantly, the proposed RCAN achieves state-of-the-art performance compared with competitive models, particularly achieving improvements of 2.6%, 1.6%, 1.4%, 1.6%, 0.3% and 0.9%, 1.7%, 1.6%, 1.2%, 0.9% in BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR metrics on both datasets, respectively.

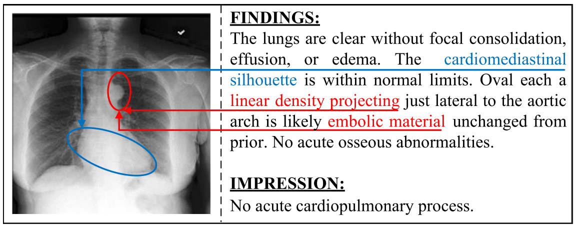
1. Introduction

Medical imaging plays a vital role in modern medical diagnosis and patient management, providing non-invasive visual information for doctors and facilitating early detection, diagnosis, and treatment of diseases (Jing, Xie, & Xing, 2018). However, converting these complex medical images into understandable and useful information remains challenging, often requiring radiologists or experts to write detailed radiology reports, which is often time-consuming, laborious, and costly (Brady, Laoide, McCarthy, & McDermott, 2012; Li, Liu, Wang, Chang, & Liang, 2023; Liu, Yin, et al., 2021). Misdiagnosis and missed diagnosis may cause serious harm to patients and even miss the best treatment opportunity (Bruno, Walker, & Abujudeh, 2015; Hou et al., 2023).

With the rapid evolution of deep learning, automated radiology report generation has become a highly focused research direction in clinical decision-making and healthcare (Yang et al., 2023; Zhou, Rueckert, & Fichtinger, 2019). The radiology report generation task aims to convert medical images into textual reports automatically. These reports typically include descriptions of anatomical structures, abnormalities, or lesions observed in the images and possible diagnoses and recommendations (Goergen et al., 2013). Automated radiology report generation has great potential to improve the efficiency and consistency of medical diagnosis, reduce the workload of doctors, and improve the quality of patient care (Biswal, Xiao, Glass, Westover, & Sun, 2020; Jing, Wang, & Xing, 2019; Li, Liang, Hu, & Xing, 2019; Pan, Yao, Li, & Mei, 2020; Syeda-Mahmood et al., 2020; Xue et al., 2018; Yuan, Liao, Luo, & Luo, 2019). Most existing radiology report generation algorithms mainly follow the image captioning model (Anderson et al., 2018; Cornia, Stefanini, Baraldi, & Cucchiara, 2020; Liu, Liu, Ren, He, & Sun, 2019). They use an encoder–decoder framework that extracts radiological visual features from images employing an encoder and

[*] Corresponding author.
E-mail addresses: houxiaodi@dlmu.edu.cn (X. Hou), xiaobo.li@dlmu.edu.cn (X. Li), lzdlmu@dlmu.edu.cn (Z. Liu), sangst@stanford.edu (S. Sang), lumingyu@dlmu.edu.cn (M. Lu), zhangyijia@dlmu.edu.cn (Y. Zhang).
[1] Equal contribution.

https://doi.org/10.1016/j.eswa.2025.126394
Received 29 May 2024; Received in revised form 6 September 2024; Accepted 2 January 2025
Available online 8 January 2025



FINDINGS:
The lungs are clear without focal consolidation, effusion, or edema. The cardiomediastinal silhouette is within normal limits. Oval each a linear density projecting just lateral to the aortic arch is likely embolic material unchanged from prior. No acute osseous abnormalities.

IMPRESSION:
No acute cardiopulmonary process.

Fig. 1. A sample represents a radiological image and its corresponding medical report. The red and blue texts correspond to clinical descriptions of different diseases, respectively.

then generate corresponding reports using a decoder. However, simply integrating image captioning methods into radiology report generation tasks may face many problems, as follows:

- **Visual and textual data bias:** The dataset contains more normal than abnormal images, resulting in imbalanced data distribution (Shin et al., 2016). And as shown in Fig. 1, the anomaly detection area is usually relatively small and only occupies a small part of the entire image (Ma et al., 2021; Vinyals, Toshev, Bengio, & Erhan, 2015). Sometimes, abnormal areas are blurry, and their structural features may not be clearly captured. In addition, the report describes the majority of text normal findings, with the imbalanced distribution of text data (Hou et al., 2023).
- **Inadequate cross-modal interaction:** Due to the dominance of normal regions throughout the image, existing models often generate similar descriptions of normal regions, neglecting the attention and description of abnormal regions (Jing et al., 2019; Li, Liang, Hu, & Xing, 2018; Xue et al., 2018). The inability to align information between different modalities may result in inconsistent reports or incomplete results (Chen, Shen, Song, & Wan, 2021). Multimodal alignment and fusion are challenging to catch clinical relatedness across diverse modalities (Ji, Sun, Dong, Wu, & Marttinen, 2022).

To capture abnormal regional features and generate accurate and continuous reports, data bias and cross-modal information alignment issues should be effectively addressed. In this paper, we propose a Recalibrated Cross-modal Alignment Network for radiology report generation with weakly supervised contrastive learning (RCAN). The RCAN develops three modules: recalibrated visual extractor module (RVE), cross-modal gated memory module (CGM), and weakly supervised contrastive learning module (WCL). RVE can alleviate visual feature bias by recalibrating abnormal features on coarse-grained visual features based on Convolutional Neural Networks (CNNs). CGM designs gated memory blocks for selectively aligning cross-modal information, achieving effective information interaction between different modalities. In addition, we introduce the WCL to embed image features into the representation space, encouraging the model to extract critical lesion structures from medical images. Overall, the paper's main contributions are as follows:

- This paper designs a recalibrated visual extractor module (RVE) to identify fine-grained visual features related to anomaly detection in radiological images to alleviate data bias issues.
- CGM achieves cross-modal information alignment by designing gating memory module to selectively memorize abnormal regions in images and text descriptions specific to anomaly findings.
- A weakly supervised contrastive learning method to maximize the difference between positive and negative cases, thereby assisting the model in learning useful medical image representations and enhancing abnormal features' detection and localization capabilities.
- Extensive experiments and analysis on the publicly available IU X-ray and MIMIC-CXR datasets validate the effectiveness of the proposed model RCAN. The experiment results show that our model RCAN achieve state-of-the-art (Sota) performance.

The section arrangement of this article is organized as follows. Section 2 reviews related studies on the radiology report generation. Section 3 describes the model framework and presents theoretical details in this paper. Section 4 shows the experimental details and conducts corresponding analysis. Section 5 summarizes the work presented in this paper. Section 6 points out the limitations of this paper and provides suggestions for future research directions.

2. Related work

The related work mainly includes three parts: Image Captioning, Radiology Report Generation and Contrastive Learning. Table 1 summaries the recent related works.

2.1. Image captioning

Image captioning is a new popular research direction following image recognition and target tracking, aiming to train computers to understand the content in images and generate a simple descriptive sentence (Anderson et al., 2018; Huang, Wang, Chen, & Wei, 2019; Liu, Ren, Liu, Wang, & Sun, 2018; Xu et al., 2015; You, Jin, Wang, Fang, & Luo, 2016). These methods mainly adopt an encoder–decoder architecture and have achieved advanced experimental performance (Liu, Ren, Liu, Lei, & Sun, 2019; Lu, Xiong, Parikh, & Socher, 2017; Rennie, Marcheret, Mroueh, Ross, & Goel, 2017). However, unlike nature images, critical information in radiological images is often concentrated in abnormal areas and contains rich details, such as small lesions, abnormal textures, etc. Due to data bias, the simple image captioning method struggles to accurately identify and describe anomalous structures in the image, which is challenging to generalize directly in radiology report generation tasks (You et al., 2021). Based on the above challenge, we design a recalibrated visual extractor module, which integrates residual blocks and two attention mechanisms to enhance the model's ability to recognize abnormal features, thereby alleviating the negative impact of data bias.

2.2. Radiology report generation

Deep learning has been successfully applied in clinical decision-making and healthcare, including providing solutions and supporting measures for generating radiology reports. During the manual radiology report generation process, radiologists or clinical experts carefully read the patient's radiology images, identify any abnormal findings, and then record them in the report (Zhang et al., 2020). This process usually requires a waste of ten minutes or even longer from the doctor, which is time-consuming and laborious (Alfarghaly, Khaled, Elkorany, Helal, & Fahmy, 2021; Jing et al., 2018). In addition, due to the similarity of medical images and the relatively small abnormal areas, the quality and efficiency of medical report generation are facing more intense challenges. Many existing works (Chen, Song, Chang, & Wan, 2020; Jing et al., 2019; Liu, Ge, & Wu, 2021; Liu, Hsu, et al., 2019) leverage deep learning algorithms to automatically generate medical reports to assist doctors in making decisions and alleviate their burden. However, due to insufficient cross-modal interaction, these models are challenging in identifying anomalous regions and descriptions associated with anomalies in radiological images. Therefore, we develop a cross-modal gated memory module by designing a gating memory mechanism that selectively stores the interaction information between abnormal visual and textual features.

2.3. Contrastive learning

Contrastive learning (Oord, Li, & Vinyals, 2018), as a machine learning method, has been successfully applied in computer vision

Table 1
The summarized of the recent related works.

Author name	Models	Key Contribution
You et al. (2016)	Image Captioning with Semantic Attention	It combines top-down and bottom-up attention mechanisms to capture global information in images and generate accurate image content.
Liu et al. (2018)	Stepwise Image-Topic Merging Network (simNet)	This model uses visual and semantic attention to integrate the extracted theme and image feature information and effectively generate comprehensive descriptions.
Chen, Song, et al. (2020)	Generating Radiology Reports via Memory-driven Transformer (R2Gen)	This model integrates the relational memory module to record critical feature information during the generation process, generating coherent and consistent radiological medical reports.
Chen et al. (2021)	Cross-modal Memory Networks for Radiology Report Generation (R2GenCMN)	It designs a cross-modal memory alignment module to achieve interaction and alignment between text and visual features.
Xue, Tan, Tan, Qin, and Xiang (2024)	Auxiliary Signal Guidance and Memory-Driven network (ASGMD)	This model introduces auxiliary signals to enhance the recognition of abnormal visual features in medical images, generating informative medical reports.
He, Fan, Wu, Xie, and Girshick (2020)	Momentum Contrast (MoCo)	This model applies contrastive learning to construct a dynamic dictionary mechanism to improve image detection and classification performance.
Ma et al. (2021)	Contrastive Attention (CA)	It compares the input medical images with normal medical images to obtain comparative information, aiming to capture visual feature representations of abnormal areas effectively.

(CV) (Chen, Fan, Girshick, & He, 2020; Chen, Kornblith, Norouzi, & Hinton, 2020) and natural language processing (NLP) (Huang, Li, Ping, & Huang, 2018; Niu, Wu, Li, & Li, 2023). The key idea of contrastive learning is to make similar data closer while dissimilar data is further away in the representation space to obtain better representation (Li, Zhang, Li, Wei, & Lu, 2023). He et al. (2020) proposed the "Momentum Contrast" method to improve image classification and detection tasks by mapping similar images to proximity regions. Gunel, Du, Conneau, and Stoyanov (2020) applied contrastive learning to pre-trained language models such as BERT to further improve the performance of NLP tasks, including text classification and named entity recognition. Tian, Shi, Li, Duan, and Xu (2018) used contrastive learning to learn the correlation between audio and video features, thereby improving the accuracy of audio-visual event localization. Recent works (Ma et al., 2021; Yan et al., 2021; Zeng, Kheir, Zeng, & Shi, 2021) have applied contrastive learning to medical image tasks and achieved competitive performance. Nevertheless, there are multiple unlabeled or incompletely labeled samples in clinical data, which is particularly important in radiology reports because fully labeled medical data is often difficult to obtain and costly. However, most previous models have overlooked this critical data resource. Inspired by the latest developments in contrastive learning, we introduce weakly supervised contrastive loss to utilize incompletely labeled data and learn feature representations via introducing weakly supervised signals, thereby improving the model's generalization ability.

3. Method

Automatically generating medical reports involves the integration of both image and textual data, constituting a multi-modal process. This task analyzes a series of input medical images denoted as $Img = \{i_1, i_2, \dots, i_{N_I}\}$ through deep learning methods. Then, critical information is extracted from the medical images to generate correspondingly described medical reports $Y = \{y_1, y_2, \dots, y_{N_R}\}$, where N_I and N_R refer to the number of radiology medical images and reports, respectively.

In this section, the proposed model RCAN is introduced. The overall structure of RCAN is shown in Fig. 2. The model mainly consists of three core parts: Recalibrated Visual Extractor, Cross-modal Gated Memory module and Weakly Supervised Contrastive Learning.

3.1. Radiology image feature input

First, given a batch of radiology medical image Img as input, following Chen et al. (2021), Chen, Song, et al. (2020), we employ the ResNet-101 as an encoder to extract the visual features from the radiology medical image. We take a series of visual features $V = \{v_1, v_2, \dots, v_{N_I}\}$ as input, where v_i is extracted through the last convolutional layer of ResNet-101. The calculation process is as follows:

$$V = \{v_1, v_2, \dots, v_{N_I}\} = f_{RIF}(Img) \tag{1}$$

where $v_i \in \mathbb{R}^{H \times W \times C}$ is the extracted visual features, where H, W and C refer to the height, width and the number of channels of the radiologic image, respectively. Where $f_{RIF}(\cdot)$ denotes the visual extractor.

Next, to retain more local feature information, the aggregated visual features $B = \{b_1, b_2, \dots, b_{N_I}\}$ are obtained through average pooling:

$$B = \{b_1, b_2, \dots, b_{N_I}\} = AvgPool(V) \tag{2}$$

where $b_i \in \mathbb{R}^{HW \times C}$, the aggregated visual features' dimension is $HW \times C$. The $AvgPool(\cdot)$ denotes the average pooling.

3.2. Recalibrated visual extractor

In order to identify fine-grained visual features related to abnormal findings in radiology images, we adopt a recalibrated visual extractor (RVE), which can enhance the ability of model RCAN to identify the key information of the extracted visual features V.

RVE introduces residual blocks to help better tune and recalibrate visual feature representations. Skip connections of residual blocks allow dynamically adjusting feature representations between channels, thereby improving the quality of feature representations. We feed the extracted visual features V into the residual block and perform the following operations:

$$K = \{k_1, k_2, \dots, k_{N_I}\} = \delta(f_r(V)) \tag{3}$$

where $k_i \in \mathbb{R}^{H \times W \times C}$, where $\delta(\cdot)$ and $f_r(\cdot)$ refers to the ReLU function and residual operation, respectively.

We employ channel attention (CA) and spatial attention (SA) to capture inter-channel and intra-channel information in medical visual features. By effectively combining these two attention mechanisms,

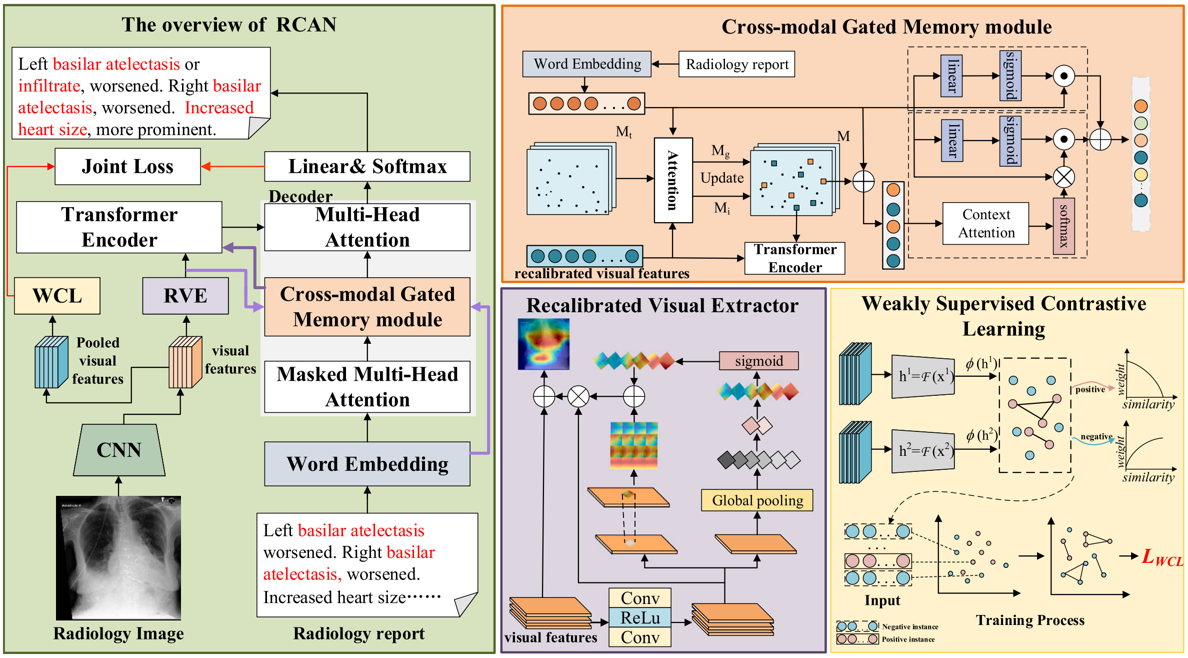


Fig. 2. The overview of the proposed model RCAN. It has three core components: Recalibrated Visual Extractor (RVE), Cross-modal Gated Memory (CGM), Weakly Supervised Contrastive Learning (WCL). The ⊕, ⊗ and ⊙ represent element-wise addition, matrix multiplication and hadamart product, respectively.

RVE can provide richer visual feature representation and better identify fine-grained visual features.

Subsequently, the channel attention receives the visual features K, which are derived from the residual block. We average the pixels within each channel to get a value that describes the content of that channel to capture global features. The following operation is as follows:

$$X = \{x_1, x_2, \dots, x_{N_I}\} = f_{GP}(K) \tag{4}$$

$$M_C = \sigma(F_{c2}((F_{c1}(X)))) \tag{5}$$

where $M_C \in \mathbb{R}^{H\times W\times C}$ refers to the visual map of channel attention, $f_{GP}(\cdot)$, $F_c(\cdot)$ and $\sigma(\cdot)$ denote the global pooling operation, the full connection and the *sigmoid* function, respectively.

Spatial attention helps the model better focus on different areas of each channel in the visual feature map. And it can help integrate information from different locations, and enhance the visual features' representation. We transmit the visual features K via the residual block into SA. Through the depth convolution operation, we can obtain the visual features $M_S \in \mathbb{R}^{H\times W\times C}$ of spatial attention maps for each channel of K.

$$M_S = Conv_{3\times 3}(K) \tag{6}$$

where $Conv_{3\times 3}(\cdot)$ denotes 3 × 3 depth convolution operation.

To make maximum use of the visual feature information within and between channels in medical images, RVE integrates channel attention and spatial attention with each other to recalibrate the visual features in images, thereby improving the quality and information extraction of medical images.

$$\hat{V} = V \oplus (\sigma(M_C \oplus M_S) \otimes K) \tag{7}$$

where ⊕ and ⊗ represent element-wise addition and matrix multiplication. The $\hat{V} = \{\hat{v_1}, \hat{v_2}, \dots, \hat{v_{N_I}}\}, \hat{v}_i \in \mathbb{R}^{H\times W\times C}$ is the recalibrated visual features.

3.3. Cross-modal gated memory module

Most models first encode the report and image separately and then use some simple methods (fully connected layer, splicing) to interactively align and generate the final report. Due to the deviation and different pattern distribution between text data and visual data, the multimodal interaction of these methods is often rigid, insufficient, and prone to adding noise information, making it difficult to achieve effective multimodal information interaction. We design a novel cross-modal gated memory module, which utilizes early interactive memory of text and image information for smooth alignment and introduces a unique gating mechanism to filter noise information and retain effective interactive features.

We randomly initialize the memory matrix $M_t = \{m_1, m_2, \dots, m_t, \dots, m_N\}$, the N denotes the number of memory vectors, and utilize attention mechanisms to interact with the word embedding in radiology reports y_g and recalibrated image visual features $\hat{v}_i$, respectively. Then, we can obtain text representation M_g and visual feature representation M_i specific to the memory matrix.

$$\begin{aligned} q_g &= y_g W_q \\ k_t &= m_t W_k \\ q_i &= \hat{v}_i W_q \end{aligned} \tag{8}$$

$$M_g = softmax(\frac{q_g k_t}{\sqrt{d_k}})m_t W_v \tag{9}$$

$$M_i = softmax(\frac{q_i k_t}{\sqrt{d_k}})m_t W_v \tag{10}$$

where W_q, W_k, and W_v denote the trainable parameters, respectively.

Because updating the memory matrix is done cyclically, we introduce U to represent the update process, as follows:

$$M = U(M_t; M_g, M_i) \tag{11}$$

A final shared memory matrix M containing rich visual and textual feature information is generated during model training and iteration.

Furthermore, to retain effective interactive features, CGM adopts a cross-modal gating. We utilize the state gate to dynamically obtain useful text embeddings in radiology reports and use the information gate to re-integrate the preserved multimodal information in the shared memory matrix and original text information. The calculation process of the state gate is as follows:

$$G_s = \sigma(W_1 \cdot y_g + b_1) \odot y_g \tag{12}$$

where G_s is the output of the state gate.

Like the state gate, the information gate refers to gating units that selectively control the input of text embedding. Furthermore, the information state also introduces a contextual attention mechanism to fuse memory matrix information and text embedding. It retains and delivers each modality's key information according to contextual relationships' needs, thereby generating richer and valuable information. The process of calculation proceeds as below:

$$G_o = \sigma(W_2 \cdot y_g + b_2) \odot y_g \tag{13}$$

$$G_i = softmax(M \cdot W_c) \otimes G_o \tag{14}$$

where W_1, W_2, W_c are the learnable matrix, and b_1, b_2 are bias. $\sigma(\cdot)$ denotes the *sigmoid* function. Where the G_i is the information gate.

We sum the outputs of the state gate G_s and the information gate G_i to obtain the final output of the cross-modal gated memory matrix G_c:

$$G_c = G_s + G_i \tag{15}$$

3.4. Weakly supervised contrastive learning

The majority of existing contrastive learning methods typically employ instance discrimination as the pretext task, treating each instance as a distinct class. However, this approach inevitably results in class collision issues that negatively impact the quality of the learned representations (Arora, Khandeparkar, Khodak, Plevrakis, & Saunshi, 2019). Weakly supervised contrastive learning utilizes graph-based approaches to identify similar samples, which employ nodes and edges to represent sample and sample similarity (connected nodes represent positive samples, negative samples represent unconnected nodes, and edges represent weakly supervised signals). The goal of loss is based on weakly supervised signals to bring the representation of similar images closer in the same space (Yan et al., 2021).

In weakly supervised contrastive learning, following SimCLR (Chen, Kornblith, et al., 2020), we utilize the aggregated visual feature $B = \{b_1, b_2, \dots, b_{N_I}\}$ obtained from Eq. (2) as input to perform data enhancement operation $A(\cdot)$, which can obtain diversity feature enhancement representation. The process is as follows:

$$\begin{aligned} b^1 &= A(b_i, \theta_1) \\ b^2 &= A(b_j, \theta_2) \end{aligned} \tag{16}$$

where θ_1 and θ_2 are random seed.

After, a CNN encoder $\mathcal{F}(\cdot)$ maps the extracted information to a low-dimensional embedding space to obtain the embedding vector, making similar images closer in the embedding space. The process is as follows:

$$\begin{aligned} h^1 &= \mathcal{F}(b^1) \\ h^2 &= \mathcal{F}(b^2) \end{aligned} \tag{17}$$

where h^1 and h^2 are embedding vector.

Then, we introduce the auxiliary projection head $\phi(\cdot)$ to find similar samples in the embedding space and generate weak labels as supervision signals. The embedding vector $P = \{p_1, p_2, \dots, p_{N_I}\}$ is extracted from the auxiliary projection head $\phi(\cdot)$ from the batch size of the samples can be written as:

$$p_i = \phi(\mathcal{F}(A(b_i, \theta))) \tag{18}$$

We input p_i into the weakly supervised contrastive loss for model training, which makes similar image representations close to each other by maximizing the loss between positive examples and minimizing the loss between negative examples. The formula is as follows:

$$\mathbb{L}_{WCL} = \sum_{i=1}^{N} \log \frac{\exp(sim(p_i, p_j)/\tau)}{\sum_{l_i \neq l_j} \exp(sim(p_i, p_j)/\tau)} \tag{19}$$

where $l_i \neq l_j$ is an indicator used to distinguish positive and negative sample contrastive categories. $sim(\cdot)$ denotes the cosine similarity between two input vectors, and τ represents the temperature parameter.

3.5. Report generator

To improve the quality of the generated radiology reports, we use the encoder–decoder architecture of Transformer as the report generator to generate radiology reports. We feed the extracted visual features fused with cross-modal memory information G_c and the recalibrated visual features $\hat{V} = \{\hat{v_1}, \hat{v_2}, \dots, \hat{v_{N_I}}\}$ into the Transformer encoder. The output is the hidden state h_i encoded from its recalibrated visual features and cross-modal geted memory for visual information. The calculation process is as follows:

$$\{h_1, h_2, \dots, h_S\} = f_e\{\hat{v_1}, \hat{v_2}, \dots, \hat{v_{N_I}}, G_c\} \tag{20}$$

where $f_e(\cdot)$ refers to the Transformer encoder.

To enhance the encoding ability of the decoder, we introduce the CGM module mentioned in Section 3.3. in the decoder to generate cross-modal memory information. This module is used together with the hidden state from the encoder and the cross-modal gated memory for text information in the decoding process, which is as follows:

$$y_t = f_d(h_1, h_2, \dots, h_S, G_c) \tag{21}$$

where $f_d(\cdot)$ denotes the Transformer decoder. Finally, to generate the final radiology report, the above process needs to be repeated until the end.

Based on the above structure, the entire process of report generation can be expressed recursively, as shown below:

$$p(Y|Img) = \prod_{t=1} p\left(y_{N_R} | y_1, y_2, \dots, y_{N_R - 1}, Img\right) \tag{22}$$

For the radiology report generation task, we use cross-entropy to calculate the loss:

$$\mathbb{L}_{CE} = -\sum_{i=1}^{N_R} \log(y_{N_R} | y_1, y_2, \dots, y_{N_R - 1}, Img) \tag{23}$$

Finally, we use the joint training of cross-entropy loss (CE) and weakly supervised contrastive loss (WCL), as follows:

$$\mathbb{L}_{loss} = \lambda \mathbb{L}_{CE} + (1 - \lambda) \mathbb{L}_{WCL} \tag{24}$$

where λ is to balance these two loss functions.

4. Experiment

In this section, we provide a detailed introduction to the experimental setup, display excellent experimental performance, and fully demonstrate the method's effectiveness.

4.1. Experiment settings

4.1.1. Datasets

We conduct experiments on the public real-world benchmark datasets IU X-ray (Shin et al., 2016)[2] and MIMIC-CXR (Johnson et al., 2019),[3] which are applied to various downstream tasks such as disease detection, image segmentation, anomaly detection.

IU X-ray: the IU X-ray dataset is a public radiology medical imaging dataset collected and released by Indiana University (IU). The dataset contains 7470 radiology medical images and 3955 radiology medical reports. Each set of medical images includes both anteroposterior and lateral views, and each report usually consists of multiple parts, including diagnostic instructions, findings, and impressions. There is no official division rule for the IU X-ray dataset. To ensure the fairness of the experiment, according to the methods of previous related work (Chen, Song, et al., 2020; Li et al., 2018), the entire dataset is randomly divided into a training set, a test set and a validation set, accounting for 70%, 20% and 10%, respectively. There is no overlap

[2] https://openi.nlm.nih.gov/

[3] https://physionet.org/content/mimic-cxr/2.0.0/

Table 2
The partition and statistics based on IU X-ray and MIMIC-CXR datasets.

Dataset	IU X-ray			MIMIC-CXR		
	Train	Val	Test	Train	Val	Test
IMAGEs	5,226	748	1,496	368,960	2,991	5,159
REPORTs	2,770	395	790	227,758	1,808	3,269
PATIENTs	2,770	395	790	64,586	500	293
AVG.LEN	37.56	36.78	33.62	53.00	53.05	66.40

between the three datasets to ensure the reliability and accuracy of the experiments.

MIMIC-CXR: the MIMIC-CXR dataset is a large-scale public chest medical imaging dataset, which includes many chest X-ray images from Beth Israel Deaconess Medical Center. The dataset contains 377,110 medical images and 227,835 radiology reports covering 65,379 patients. To ensure the comparability of experimental results, we group the dataset according to the official division rules. According to the 7:1:2, the dataset is divided into a training set, a validation set and a test set.

In Table 2, we divide these two datasets into training, test, and validation sets and count the number of images, reports, patients, and average word length of the report.

4.1.2. Implementation detail

Following previous works (Chen, Song, et al., 2020; Li et al., 2018), we utilize two medical images of the patient as input for the IU X-ray report and one for the MIMIC-CXR report. To extract visual features, we use the ResNet-101 (He, Zhang, Ren, & Sun, 2016) pre-trained model on ImageNet as a visual encoder to extract visual features of the image, with the dimension of each feature being $d = 2048$. All radiology medical images are cropped to 224 × 224. For the datasets MIMIC-CXR and IU X-ray, we set the length of the generated report to 100 and 60, respectively.

We use a Transformer model with 3 layers and 8 attention heads for the report generator, with the hidden state set to 512 dimensions and randomly initialized. The model RCAN is trained under CE and WCL employing Adam (Kingma & Ba, 2015) as the optimizer. The learning rate of the visual extractor is $5e-4$, and $1e-4$ for other parameters. We train the IU X-ray and MIMIC-CXR datasets for 100 and 30 epoches. We decrease the learning rate by 0.8 at the end of each epoch. The beam search strategy is used with a beam size of 3. Follow the settings from previous work (Hou et al., 2023; Yang, Zou, Wu, Xu, & Fan, 2022), in weakly supervised contrastive learning, hyperparameter $\tau \in \{0.05, 0.07, 0.1, 0.15, 0.2\}$, $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

We train our model on the training set, optimize and adjust the model's hyperparameters on the validation set. We evaluate the proposed model RCAN, choose the generated model with the highest evaluation BLEU-4 on the validation set and employ the model to generate medical reports on the test set.

4.1.3. Evaluation metrics

To ensure a fair comparison with existing models, we employ common natural language generation (NLG) metrics to gauge the quality of generated radiology reports as an integral part of the report generation process. Specifically, we employ BLEU-n (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004) metrics, which gauge the correspondence between the generated reports and the ground truth reports. BLEU-n measures similarity by comparing the generated report with the matching of n consecutive phrases (n-grams) in the ground truth. The METEOR evaluation index considers word level similarity and includes sentence level matching, grammatical structure matching, and attention to word order correctness, which enables it to more accurately evaluate the fluency and structure of the generated text and better reflect its quality. ROUGE-L measures similarity by comparing the longest common subsequence (LCS) between the generated summary text and the reference summary text, which helps to capture important components of text structure and content. Therefore, ROUGE-L can comprehensively consider the accuracy of generating reports.

4.1.4. Comparison methods

- **HRGR-Agent** (Li et al., 2018): it utilizes reinforcement learning to integrate retrieval-based and learning-based generation, generating diverse and robust reports by learning different word-level and sentence-level features.
- **COATT** (Jing et al., 2018): it uses a co-attention mechanism to learn the features of image and text simultaneously and designs a hierarchical LSTM to model reports over long distances.
- **CMAS-RL** (Jing et al., 2019): it trains a new system, a cooperative multi-agent system, using reinforcement algorithms to utilize three agents (planner, abnormality writer, and normality writer) to describe detected anomalies and generate meaningful reports.
- **R2Gen** (Chen, Song, et al., 2020): it applies the designed relational memory to the Transformer architecture to capture attention mapping between medical images and report text. Particularly, it reports experimental results for the first time on the MIMIC-CXR dataset.
- **CMCL** (Liu, Ge, & Wu, 2021): it imitates the radiologists learning to write reports (the difficulty of learning reports ranges from easy to challenging), allowing the model to gradually learn simple and complex samples to alleviate data bias.
- **R2GenCMN** (Chen et al., 2021): it emphasizes the importance of cross-modal mapping for information enhancement, which designs a cross-modal memory matrix to perform memory queries and responses on image and text information.
- **ASGMD** (Xue et al., 2024): it regards medical subject entities as auxiliary signals and designs a memory-driven network to alleviate data bias issues.
- **PPKED** (Liu, Wu, et al., 2021): it integrates medical knowledge retrieved from radiology reports and domain medical knowledge maps into the model to alleviate visual and textual data bias, improving its ability to describe rare anomalies.
- **Clinical-BERT** (Yan & Pei, 2022): it pre-trains language models on the MIMIC-CXR dataset and combines pre-trained language models with medical domain knowledge to serve downstream medical report generation tasks.

4.2. Experiment results

Due to the different distribution of datasets used by different models, we compare multiple datasets based on different models for experiments to verify the effectiveness of our proposed method. Among them, HRGR-Agent (Li et al., 2018), COATT (Jing et al., 2018), CMAS-RL (Jing et al., 2019), R2Gen (Chen, Song, et al., 2020), CMCL (Liu, Ge, & Wu, 2021), R2GenCMN (Chen et al., 2021), PPKED (Liu, Wu, et al., 2021), ASGMD (Xue et al., 2024), Clinical-BERT (Yan & Pei, 2022) are based on the IU X-ray dataset, and R2Gen (Chen, Song, et al., 2020), CMCL (Liu, Ge, & Wu, 2021), R2GenCMN (Chen et al., 2021), ASGMD (Xue et al., 2024), Clinical-BERT (Yan & Pei, 2022) are based on the MIMIC-CXR dataset.

In Table 3, our proposed RCAN achieves SOTA performance in all indicators tested on the IU X-ray dataset and surpasses other models in most indicators on the MIMIC-CXR dataset. R2GenCMN is a classic baseline model that designs a memory network to facilitate cross-modal information exchange, but it struggles to capture fine-grained abnormal visual feature information. Clinical-BERT is a competitive model that attempts to learn medical domain knowledge and uses attention for modal alignment during model pre-training, neglecting concentration on abnormal regions and resulting in limited performance. RCAN exceeds Clinical-BERT by 2.6%, 1.6%, 1.4%, 1.6%, 0.3%,

Table 3
The performance of our proposed model RCAN compared with previous models on the publicly available datasets IU X-ray and MIMIC-CXR. The best performance is represented in bold font.

Dataset	MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-ray	NIC* (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	–	0.306
	ADAATT (Lu et al., 2017)	0.220	0.127	0.089	0.068	–	0.308
	Att2IN* (Rennie et al., 2017)	0.224	0.129	0.089	0.068	–	0.308
	SA&T* (Xu et al., 2015)	0.339	0.251	0.168	0.118	–	0.323
	HRGR-Agent (Li et al., 2018)	0.216	0.124	0.087	0.066	–	0.322
	COATT (Jing et al., 2018)	0.455	0.288	0.205	0.154	–	0.369
	CMAS-RL (Jing et al., 2019)	0.464	0.301	0.210	0.154	–	0.362
	R2Gen (Chen, Song, et al., 2020)	0.470	0.304	0.219	0.165	0.187	0.371
	CMCL (Liu, Ge, & Wu, 2021)	0.473	0.305	0.217	0.162	0.186	0.378
	R2GenCMN (Chen et al., 2021)	0.475	0.309	0.222	0.170	0.191	0.375
	PPKED (Liu, Wu, Ge, Fan, & Zou, 2021)	0.483	0.315	0.224	0.168	0.190	0.376
	Clinical-BERT (Yan & Pei, 2022)	0.495	0.330	0.231	0.170	0.209	0.376
	ASGMD (Xue et al., 2024)	0.489	0.326	0.232	0.173	0.206	0.397
	RCAN	**0.521**	**0.346**	**0.245**	**0.186**	**0.212**	**0.399**
MIMIC -CXR	NIC* (Vinyals et al., 2015)	0.299	0.184	0.121	0.084	0.124	0.263
	ADAATT (Lu et al., 2017)	0.299	0.185	0.124	0.088	0.118	0.266
	Att2IN* (Rennie et al., 2017)	0.325	0.203	0.136	0.096	0.134	0.276
	Up-Down (Anderson et al., 2018)	0.317	0.195	0.130	0.092	0.128	0.267
	CMCL (Liu, Ge, & Wu, 2021)	0.344	0.217	0.140	0.097	0.133	0.281
	R2Gen (Chen, Song, et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277
	R2GenCMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.278
	PPKED (Liu, Wu, et al., 2021)	0.360	0.224	0.149	0.106	0.149	0.284
	Clinical-BERT (Yan & Pei, 2022)	0.383	0.230	0.151	0.106	0.144	0.270
	ASGMD (Xue et al., 2024)	0.372	0.233	0.154	0.112	0.152	**0.286**
	RCAN	**0.392**	**0.247**	**0.167**	**0.118**	**0.153**	0.275

Table 4
The performance of ablation experiments on various components in the proposed model RCAN, with the best performance represented in bold font.

Dataset	RVE	CGM	WCL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
IU X-ray	✓			0.496	0.325	0.225	0.162	0.209	0.376
		✓		0.496	0.324	0.225	0.160	0.205	0.382
			✓	0.485	0.309	0.218	0.158	0.210	0.377
	✓	✓		0.504	0.327	0.235	0.176	0.211	0.383
	✓		✓	0.490	0.311	0.219	0.158	0.210	0.376
		✓	✓	0.502	0.337	0.236	0.170	0.209	0.364
	✓	✓	✓	**0.521**	**0.346**	**0.245**	**0.186**	**0.212**	**0.399**
MIMIC -CXR	✓			0.344	0.211	0.140	0.100	0.133	**0.275**
		✓		**0.396**	0.228	0.145	0.098	0.143	0.251
			✓	0.305	0.189	0.125	0.087	0.108	0.256
	✓	✓		0.372	0.223	0.148	0.103	0.137	0.275
	✓		✓	0.303	0.188	0.124	0.087	0.109	0.255
		✓	✓	0.372	0.232	0.157	0.112	0.140	0.276
	✓	✓	✓	0.392	**0.247**	**0.167**	**0.118**	**0.153**	**0.275**

and 2.3% in terms of BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE-L indicators, respectively. Compared to the recently proposed ASGMD, it introduces auxiliary signals to focus on abnormal areas in the image and achieves good performance. Compared with ASGMD, we introduce weakly supervised contrastive loss to fully utilize unlabeled and incompletely labeled data, further enhancing the model's ability to extract anomalous features. Specifically, we comprehensively surpass ASGMD on the IU X-ray dataset, particularly in achieving significant improvements in BLEU-1, BLEU-2, BLEU-3, and BLEU-4 metrics, which are 3.2%, 2.0%, 1.3%, and 1.3%, respectively.

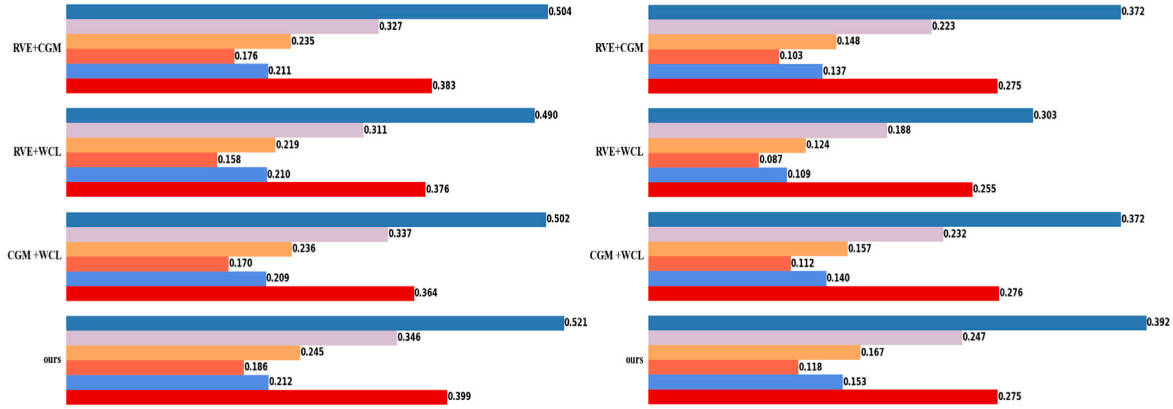
4.3. Module ablation study

The proposed model RCAN achieves the highest BLEU-4 score on MIMIC-CXR and IU X-ray datasets, which shows that our model can generate more accurate and richer words and phrases. At the same time, in these two datasets, the METEOR indicator also obtained the highest score, indicating that our proposed model can generate fluent radiology medical imaging reports with clinical medical terminology. In addition, the ROUGE-L index is better than other models, which shows that the medical reports generated by the RCAN model are closer to the ground truth and have higher clinical accuracy. This demonstrates that our model performs well in identifying abnormal regions, capturing long sequence-level sentences, and generating clinical report consistency and accuracy.

The module effectiveness is evaluated via a series of ablation experiments performed on the IU X-ray and MIMIC-CXR datasets. The RCAN model contains three core modules: RVE, CGM, WCL, and we evaluate the three module's effectiveness using BLEU-n, METEOR, and ROUGE-L metrics.

- **RVE:** the **R**ecalibrated **V**isual **E**xtractor (RVE) aims to identify fine-grained visual features related to anomaly detection in radiological images.
- **CGM:** the **C**ross-modal **G**ated **M**emory module (CGM) aims to achieve cross-modal information alignment by designing gating memory module.
- **WCL:** the **W**eakly supervised **C**ontrastive **L**earning (WCL) aims to assist the model in learning useful medical image representations and enhance abnormal features' detection capability.

We compare and visually analyze the performance of different indicators of the seven methods. In Table 4, we perform the ablation experiment settings of adding only one module, removing only one module, and combining all modules. The rich experimental results show that the experimental performance drops significantly when a



(a) IU X-Ray

(b) MIMIC-CXR

Fig. 3. The module ablation study based on the IU X-ray and MIMIC-CXR datasets. Six colors from top to bottom represent six indicators: BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L.

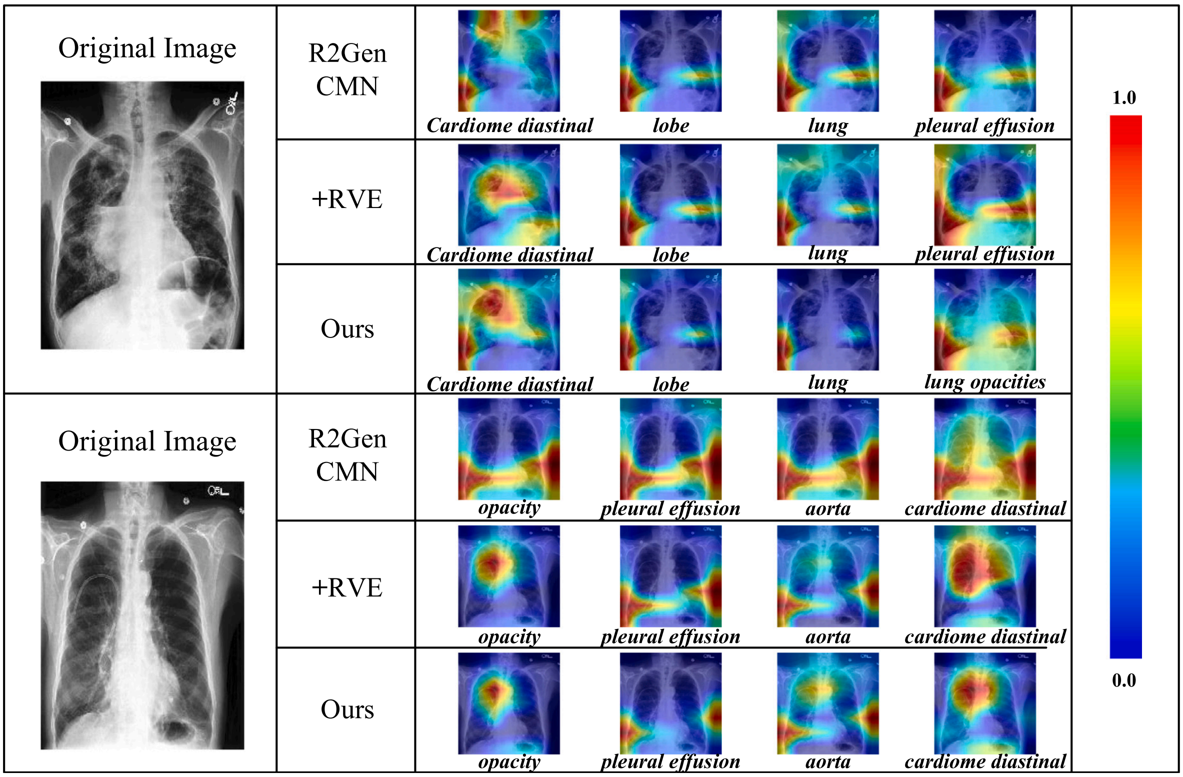


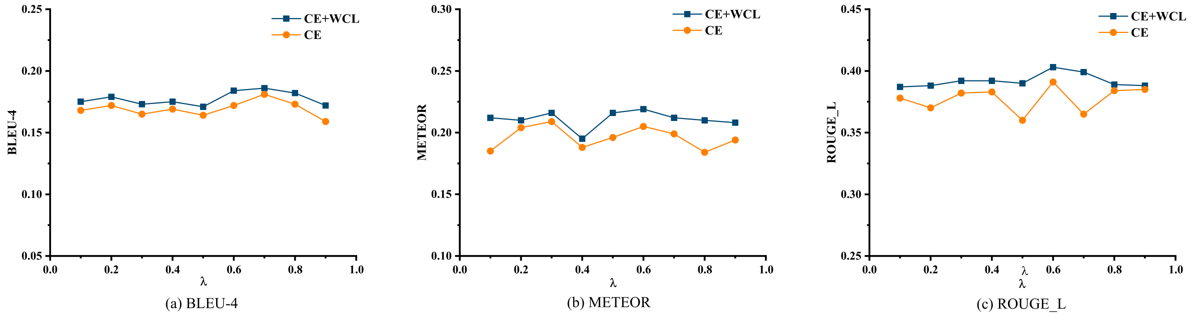
Fig. 4. The visualization of the R2GenCMN, R2GenCMN + Recalibrated Visual Extractor (RVE) and ours. In visualized medical images, redder colors represent higher weights, and bluer colors represent lower weights.

specific module is removed, or only a particular module is used. The overall performance is optimal when all three modules are combined to form our RCAN. In Fig. 3, we carry out a visual analysis of only removing a certain module. It can be seen intuitively and clearly that when our RCAN removes a certain module on different datasets, the experimental performance decreases, demonstrating the effectiveness and compatibility of the three modules.

4.4. The effect of RVE

The recalibrated visual extractor captures fine-grained visual features of abnormal findings in medical images, providing key information for subsequent report generation. To confirm the efficacy of this module, we randomly select two radiology medical images in the MIMIC-CXR dataset, apply them to R2GenCMN, RVE, and our proposed model RCAN, respectively, and perform attention view visualization.

In Fig. 4, we show each word in each report generated by different models and its corresponding visualization results of the radiology medical imaging attention map. As shown in Fig. 4, compared with the R2GenCMN model, the recalibrated visual extractor is more capable of capturing diseases and organ findings in medical images at a finer granularity, further enhancing the overall model RCAN's attention to specific locations in medical images, thereby improving medical image report quality.



(a) BLEU-4 (b) METEOR (c) ROUGE_L

Fig. 5. Comparing the performance of cross-entropy loss (CE) and cross-entropy loss combined with weakly supervised contrastive loss (CE+WCL) in the proposed model RCAN with different λ on the IU X-ray dataset. The blue and orange lines represent CE+WCL and CE, respectively.

4.5. The effect of WCL

Radiology report generation is a task of image-to-text generation. In the model training stage, we use a cross-entropy loss combined with a weakly supervised contrastive learning loss (CE+WCL) to train the model. The CE is used to evaluate the difference between the radiology report generated by the RCAN model and the ground truth, thereby generating more accurate and high-quality medical text reports. The WCL can assist the model to learn the representation of medical images, which can capture better the characteristics and structure of lesions in medical images. In this process, the hyperparameter λ in Eq. (21) plays a role in balancing WCL and cross-entropy (CE).

In Fig. 5, we conduct a comparative experiment comparing the training results using CE and CE combined with WCL (CE+WCL). We use three evaluation indicators: BLEU-4, METROR, and ROUGE_L. These indicators evaluate the quality of the generated reports. In Fig. 5, the horizontal axis represents the value of λ, and the vertical axis represents the value of the evaluation index. The blue line represents the model's performance when training with two loss functions simultaneously, the orange line represents the case where the model only uses CE for training. In Fig. 5, we can observe the following phenomena:

- The overall performance of the evaluation indicators BLEU-4, METEOR, and ROUGE_L is lower than when only employed the CE to train the model. In comparison, when the model RCAN combines the WCL for training, these evaluations of the indicator's performance improved. This indicates that the WCL is crucial in the radiology report generation task. The WCL helps capture the lesion characteristics and findings in the image, which can improve RCAN's ability to identify visual features of medical images and alleviate data bias problems; introducing WCL helps RCAN generate text descriptions related to medical image finding, thereby generating accurate, coherent, and high-quality radiology reports.
- When $\lambda = 0.7$, the value of BLEU-4 reaches its highest, and when $\lambda = 0.6$, the values of METEOR and ROUGE_L are the highest, but when the value of λ is too low or too high, the performance of the model is poor. This indicates the importance of adjusting λ appropriately to balance the weight between cross-entropy loss and weakly supervised contrastive loss. These experimental results demonstrate that appropriate adjustments can effectively assist the model RCAN's performance, which improves the accuracy and consistency of generated reports, making them more in line with medical standards. Therefore, the introduction of WCL and the selection of appropriate λ are crucial for optimizing the model's performance in the radiology report generation task, which helps to generate more accurate, consistent, and high-quality radiological reports.

Table 5
The performance of different τ experiment of the model on IU X-ray dataset.

τ	BL-1	BL-2	BL-3	BL-4	MET	RG_L
0.05	0.492	0.315	0.227	0.172	0.192	0.387
0.07	0.505	0.328	0.226	0.165	0.208	0.390
0.1	**0.521**	**0.346**	**0.245**	**0.186**	**0.212**	**0.399**
0.15	0.491	0.313	0.221	0.169	0.210	0.378
0.2	0.498	0.315	0.214	0.165	0.207	0.381

4.6. The impact of the parameter τ

Previous radiology report generation methods (Liu, Yin, et al., 2021; Zeng et al., 2021) that applied contrastive learning overlook using unlabeled or incompletely labeled samples in the data. Weakly supervised learning allows for the use of unlabeled or incompletely labeled data for training, thereby enhancing attention to clinically relevant information and improving the robustness and stability of the model. In weakly supervised contrastive learning, τ is a crucial hyperparameter used to fine-tune the sensitivity of the WCL. The choice of temperature parameters τ directly affects the smoothness of the distance distribution and the contrast between positive and negative samples. By adequately adjusting this parameter, we can better capture the internal structure of radiology medical images, thereby improving the recognition performance of abnormal visual features.

We conduct experiments using the above evaluation metrics to evaluate model performance under different temperature parameters τ. In Table 5, all evaluation indicators achieve the best results when $\tau = 0.1$. This result shows that when τ is too low, the distance between similar samples increases, resulting in too high contrast, making it difficult to distinguish between positive and negative samples. When τ is too high, the distance decreases, resulting in a gap between positive and negative samples. It is difficult to distinguish and capture the slight differences between medical images. This can also result in the generated medical reports being overly descriptive of normal findings and failing to generate descriptions of abnormal findings, thus inducing visual and textual biases.

Overall, a reasonable selection of τ is crucial in applying the WCL in radiology report generation. The temperature parameter affects the stability and convergence speed of the model during training by adjusting the smoothness of the loss function. Smaller τ may make the training process more unstable, but they help the model capture finer differences. Larger τ may make the training process smoother but weaken the ability to distinguish abnormal areas.

4.7. Case study

To prove the efficacy of the model RCAN proposed in this paper, we randomly select two groups of generated medical reports for analysis. As shown in Fig. 6, we compare the medical reports we generated with the ground truth, Transformer method, and R2GenCMN method.

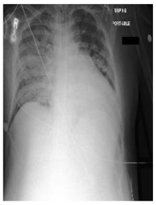
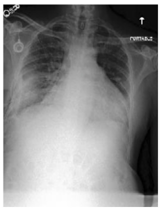
Images	Ground Truth	Transformer	R2GenCMN	Ours
	The ET tube is 3.5 cm above the carina. The NG tube tip is off the film, at least in the stomach. Right IJ Cordis tip is in the proximal SVC. The heart size is moderately enlarged. There is ill-defined vasculature and alveolar infiltrate, right greater than left. This is markedly increased compared to the film from two hours prior and likely represents fluid overload.	The NG tube tip is off the film, the cardiac grossly unremarkable. The heart size is big. Unclear fluid appears in the alveolar	There is unclear fluid in the alveolar. The tip of the NG tube has detached from the film. The tip of the right IJ Cordis is located near the superior vena cava. Moderate increase in heart size. There is unclear vascular system and alveolar infiltration.	The endotracheal tube is terminating approximately 3.5 cm above the level of the carina . The NG tube tip is coursing below the level of the diaphragm into the expected location of the stomach. Right IJ Cordis tip is in the proximal SVC, heart size is moderately enlarged. the cardiac and mediastinal silhouettes are grossly unremarkable . Vascularity, infiltration of the alveolar. right greater than left.
	as compared to the previous radiograph the patient has received a right chest tube . the chest tube is in correct position . there is a substantial decrease in extent of the pre-existing right pleural effusion . no pneumothorax is seen . minimal air collection in the soft tissues at the site of drain insertion . unchanged moderate cardiomegaly. unchanged normal appearance of the left lung . right port-a-cath in situ .	The chest tube is in the correct position. The range of right pleural effusion present is reduced. No pleural effusion, left lung shape unchanged.	The chest tube is in the correct position. There is already a significant reduction in pleural effusion. Presence of cardiac hypertrophy, normal left lung shape and appearance, catheter in correct position.	right-sided chest tupe-a-cath tip terminates in correct position . moderate enlargement of the cardiac silhouette is redemonstrated . the left lung, mediastinal, hilar contours are unchanged . pulmonary vasculature is not engorged . patchy opacities within the lung bases likely reflect areas of atelectasis . no pneumothorax presents .there is a substantial decrease in extent of the pre-existing right pleural effusion

Fig. 6. Examples of the generated reports. Our results are compared with the ground truth, the Transformer method and R2GenCMN method.

The red fonts represent organs and medical equipment, and the blue fonts represent basic facts and disease descriptions. The similar descriptions for the input medical images in the report are marked in the same color.

Fig. 6 shows that the reports generated by our RCAN model are more comprehensive and closer to the description of ground truth compared to the Transformer method and R2GenCMN method. In the first example, our model successfully and accurately presents key information such as *"NG tube tip"*, *"stomach"* and *"vascularity"*, while the other two methods do not provide detailed descriptions of *"heart"* and *"volatility"*. This indicates that our proposed model can capture small features in images, providing more comprehensive support for the complete presentation of images. In addition, RCAN can generate long text reports that meet clinical standards, such as "The endotracheal tube is terminating approximately 3.5 cm above the level of the carina", demonstrating its accuracy and availability in generating clinical reports. In the second example, the report generated by RCAN is more relative to the real radiology report, and it is easier to observe that compared to the other two methods, RCAN's description of *"pleural effusion"* and *"chest tube"* is more detailed and comprehensive.

It can be seen from the above results that the report generated by the Transformer cannot fully find and convey the findings in the medical images, and the description of details by R2GenCMN is not meticulous enough. The RCAN model proposed in this paper can recognize radiological medical images and generate coherent and clinically meaningful medical reports.

However, our model still needs further improvement. In the second example, RCAN ignores the prediction and description of *"soft tissues"*, which requires future work to improve our model performance. In future work, we will improve the model's ability to recognize and generate anomalies in medical images, enabling it to better capture and describe subtle features. In addition, we will explore the interaction and fusion methods between multimodal data to integrate text and image information better, aiming to make the connections between different data types closer, thereby improving the consistency and accuracy of generating medical report information.

5. Conclusions

In this paper, we propose a Recalibrated Cross-modal Alignment Network for radiology report generation with weakly supervised contrastive learning (RCAN), which can alleviate data bias and align cross-modal information. The RCAN develops three modules: recalibrated visual extractor module (RVE), cross-modal gated memory module (CGM), and weakly supervised contrastive learning module (WCL), to capture abnormal regional features and generate accurate and continuous reports. The RVE integrates residual blocks and two attention mechanisms to enhance the model's ability to recognize abnormal features, thereby alleviating the negative impact of data bias. The CGM designs a gating memory mechanism that selectively stores the interaction information between abnormal visual and textual features. Moreover, the WCL learns feature representations through weakly supervised methods to enhance attention to anomalous features.

Overall, we conduct extensive experimental verification on two real-world datasets: IU X-ray and MIMIC-CXR. The results show that our model RCAN exceeds all competitive models in most metrics, particularly achieving improvements of 2.6%, 1.6%, 1.4%, 1.6%, 0.3% and 0.9%, 1.7%, 1.6%, 1.2%, 0.9% in BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR metrics on both datasets, respectively. The ablation study also demonstrates the effectiveness and compatibility of all core modules in the paper. Furthermore, we visualize and compare the reports generated by different models, and the results show that our model can more accurately identify abnormal features in images and generate more continuous medical reports.

6. Limitations and future work

Although the recalibrated cross-modal alignment network we designed can significantly improve radiology report generation performance, it cannot explicitly generate the process and lacks interpretability. In addition, due to the difficulty in obtaining medical data, the model's effectiveness has not been validated on different clinical datasets, which may limit the model's generalization. In the future, we plan to explore the usability of the model in various clinical environments to enhance its robustness and generalization. Meanwhile, further improving the transparency and interpretability of report generation is also essential for future work.

CRediT authorship contribution statement

Xiaodi Hou: Designed the method and experiments, Performed the experiments and analyzed the results, Writing – original draft. **Xiaobo Li:** Performed the experiments and analyzed the results, Writing – orginal draft. **Zhi Liu:** Provided suggestions and feedback. **Shengtian Sang:** Provided suggestions and feedback. **Mingyu Lu:** Provided suggestions and feedback. **Yijia Zhang:** Designed the method and experiments, Provided suggestions and feedback, Funding acquisition.

Code availability

The code is available at: https://github.com/Eleanorhxd/RCAN.git.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by grant from the Natural Science Foundation of China (No. 62072070).

Data availability

Data will be made available on request.

References

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, *24*, Article 100557.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077–6086).

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., & Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. arXiv preprint arXiv:1902.09229.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Biswal, S., Xiao, C., Glass, L. M., Westover, B., & Sun, J. (2020). Clara: clinical report auto-completion. In *Proceedings of the web conference 2020* (pp. 541–550).

Brady, A., Laoide, R. Ó., McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster Medical Journal*, *81*(1), 3.

Bruno, M. A., Walker, E. A., & Abujudeh, H. H. (2015). Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, *35*(6), 1668–1676.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, Z., Shen, Y., Song, Y., & Wan, X. (2021). Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 5904–5914). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.459.

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 1439–1449). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.112.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10578–10587).

Goergen, S. K., Pool, F. J., Turner, T. J., Grimm, J. E., Appleyard, M. N., Crock, C., et al. (2013). Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *Journal of Medical Imaging and Radiation Oncology*, *57*(1), 1–7.

Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Hou, X., Liu, Z., Li, X., Li, X., Sang, S., & Zhang, Y. (2023). MKCL: Medical knowledge with contrastive learning model for radiology report generation. *Journal of Biomedical Informatics*, *146*, Article 104496. http://dx.doi.org/10.1016/j.jbi.2023.104496, URL https://www.sciencedirect.com/science/article/pii/S1532046423002174.

Huang, J., Li, Y., Ping, W., & Huang, L. (2018). Large margin neural language model. arXiv preprint arXiv:1808.08987.

Huang, L., Wang, W., Chen, J., & Wei, X.-Y. (2019). Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4634–4643).

Ji, S., Sun, W., Dong, H., Wu, H., & Marttinen, P. (2022). A unified review of deep learning for automated medical coding. arXiv preprint arXiv:2201.02797.

Jing, B., Wang, Z., & Xing, E. (2019). Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6570–6580). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1657.

Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2577–2586). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1240.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings*.

Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in Neural Information Processing Systems*, *31*.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 33*, (no. 01), (pp. 6666–6673).

Li, M., Liu, R., Wang, F., Chang, X., & Liang, X. (2023). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, *26*(1), 253–270.

Li, X., Zhang, Y., Li, X., Wei, H., & Lu, M. (2023). DGCL: Distance-wise and graph contrastive learning for medication recommendation. *Journal of Biomedical Informatics*, *139*, Article 104301.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Liu, F., Ge, S., & Wu, X. (2021). Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 3001–3012). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.234.

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., et al. (2019). Clinically accurate chest X-ray report generation. In *Machine learning for healthcare conference* (pp. 249–269). PMLR.

Liu, F., Liu, Y., Ren, X., He, X., & Sun, X. (2019). Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, *32*.

Liu, F., Ren, X., Liu, Y., Lei, K., & Sun, X. (2019). Exploring and distilling cross-modal information for image captioning. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 5095–5101). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2019/708.

Liu, F., Ren, X., Liu, Y., Wang, H., & Sun, X. (2018). Simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 137–149). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1013.

Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13753–13762).

Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., & Sun, X. (2021). Contrastive attention for automatic chest X-ray report generation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 269–280). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-acl.23.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 375–383).

Ma, X., Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., et al. (2021). Contrastive attention for automatic chest X-ray report generation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 269–280). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-acl.23.

Niu, K., Wu, Y., Li, Y., & Li, M. (2023). Retrieve and rerank for automated ICD coding via contrastive learning. *Journal of Biomedical Informatics*, *143*, Article 104396.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10971–10980).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 7008–7024).

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., & Summers, R. M. (2016). Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2497–2506).

Syeda-Mahmood, T., Wong, K. C., Gur, Y., Wu, J. T., Jadhav, A., Kashyap, S., et al. (2020). Chest X-ray report generation through fine-grained label learning. In *Medical image computing and computer assisted intervention–MICCAI 2020: 23rd international conference, Lima, Peru, October 4–8, 2020, proceedings, part II 23* (pp. 561–571). Springer.

Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C. (2018). Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision* (pp. 247–263).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057). PMLR.

Xue, Y., Tan, Y., Tan, L., Qin, J., & Xiang, X. (2024). Generating radiology reports via auxiliary signal guidance and a memory-driven network. *Expert Systems with Applications*, *237*, Article 121260. http://dx.doi.org/10.1016/j.eswa.2023.121260, URL https://www.sciencedirect.com/science/article/pii/S0957417423017621.

Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., et al. (2018). Multimodal recurrent model with attention for automated radiology report generation. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, proceedings, part I* (pp. 457–466). Springer.

Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., et al. (2021). Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4009–4015). Association for Computational Linguistics.

Yan, B., & Pei, M. (2022). Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 36*, (no. 3), (pp. 2982–2990).

Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., & Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, *86*, Article 102798.

Yang, C., Zou, J., Wu, J., Xu, H., & Fan, S. (2022). Supervised contrastive learning for recommendation. *Knowledge-Based Systems*, *258*, Article 109973.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651–4659).

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., & Wu, X. (2021). Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part III 24* (pp. 72–82). Springer.

Yuan, J., Liao, H., Luo, R., & Luo, J. (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part VI 22* (pp. 721–729). Springer.

Zeng, D., Kheir, J. N., Zeng, P., & Shi, Y. (2021). Contrastive learning with temporal correlated medical images: A case study using lung segmentation in chest X-Rays. In *2021 IEEE/ACM international conference on computer aided design* (pp. 1–7). IEEE.

Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., & Xu, D. (2020). When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (pp. 12910–12917).

Zhou, S. K., Rueckert, D., & Fichtinger, G. (2019). *Handbook of medical image computing and computer assisted intervention*. Academic Press.