



# Generating radiology reports via auxiliary signal guidance and a memory-driven network

Youyuan Xue<sup>a</sup>, Yun Tan<sup>a,\*</sup>, Ling Tan<sup>b</sup>, Jiaohua Qin<sup>a</sup>, Xuyu Xiang<sup>a</sup>

<sup>a</sup> College of Computer Science and Information Technology, Central South University of Forestry and Technology, Changsha 410004, China

<sup>b</sup> The Second Xiangya Hospital of Central South University, Changsha, 410011, China

## ARTICLE INFO

Dataset link: <https://github.com/shangchengLu/ASGMDN>

### Keywords:

Radiology reports generation  
Auxiliary signal  
Attention mechanism  
Memory mechanism

## ABSTRACT

Automatically generating medical image reports is a gratifying task. For doctors, it can reduce the heavy burden of writing reports, and for patients, it can reduce the waiting time for reports; it can also avoid misdiagnosis and missed diagnoses caused by human factors. However, this task still faces enormous challenges due to the problem of visual and textual data bias and the complex relationships among the components of medical reports. To this end, in this work, we propose an auxiliary signal guidance and memory-driven (ASGMD) network that can be used to generate medical reports automatically. It includes three modules: an Auxiliary Signal Guidance Module (ASG), a text sequential attention mechanism (TSAM) module, and a Memory Mechanism-Driven Decoding Module (MMDD). Given a medical image of a patient, radiologists usually focus on the abnormal area first, then browse the global information included in the image and write a corresponding report. Similar to the above working mode, the ASG module enhances the features of the abnormal areas of medical images by introducing auxiliary signals that alleviate the problem of visual data bias. We design a novel TSAM module that explores the consistency of medical report context and enhances essential medical information in reports to reduce textual data bias. Finally, the MMDD module integrates visual and textual knowledge to achieve dynamic decoding and generate a final report. The experimental results show that the proposed method outperforms state-of-the-art models on various evaluation metrics on the two public datasets, IU-Xray and MIMIC-CXR. To make our results reproducible, our code has been released at <https://github.com/shangchengLu/ASGMDN>.

## 1. Introduction

With the development of high-precision medical equipment and medical imaging technology, medical personnel can more directly examine abnormalities inside a patient's body through imaging data and thus make more accurate diagnoses. However, writing an imaging report is laborious and error-prone for inexperienced radiologists and time-consuming and tedious for experienced clinicians. At the same time, the objectivity of medical reports written by medical staff is affected by factors such as emotion and mental state. In addition, patients often must wait for an extended period to obtain the corresponding medical image report, and this significantly delays the patient's diagnosis time. Therefore, it is critical to develop a method for acquiring objective and precise medical imaging reports quickly and automatically.

Most of the existing works in this area of research, such as Ren et al. (2015), Rennie et al. (2017), and Dognin et al. (2019), follow the paradigm of image captioning and utilize encoder-decoder

frameworks, e.g., CNN-RNN. Jing et al. (2018) proposed the first fully automatic model for generating medical image reports based on an encoder-decoder framework. In the encoder part of the model, CNN is responsible for extracting image features. The decoder part consists of Sentence LSTM and Word LSTM, which generate paragraphs and words. Although this method achieves excellent experimental results, it lacks contextual consistency and generates many duplicate reports due to its hierarchical LSTMs. Xue et al. (2018) proposed constructing an attention input to guide the generation of subsequent sentences by combining the encoding of images and the generated sentences, thereby maintaining coherence of the generated sentences. However, the model cannot generate abnormal reports due to data bias (it yields a severely imbalanced data distribution). Some recent work has addressed the problem of data bias in medical report generation. For example, Harzig et al. (2019) proposed the use of dual LSTM to generate standard and abnormal report information separately, increasing the variability of the generated text. At the same time, attention mechanisms, which

\* Corresponding author.

E-mail addresses: [lushangcheng7@gmail.com](mailto:lushangcheng7@gmail.com) (Y. Xue), [tantanyun@hotmail.com](mailto:tantanyun@hotmail.com) (Y. Tan), [dr.tanling@csu.edu.cn](mailto:dr.tanling@csu.edu.cn) (L. Tan), [qinjiaohua@163.com](mailto:qinjiaohua@163.com) (J. Qin), [xyuxiang@163.com](mailto:xyuxiang@163.com) (X. Xiang).

<https://doi.org/10.1016/j.eswa.2023.121260>

Received 22 June 2022; Received in revised form 14 August 2023; Accepted 18 August 2023

Available online 9 September 2023

0957-4174/© 2023 Elsevier Ltd. All rights reserved.

continue to be developed, have been widely used in various image description deep learning models. Inspired by [Chen et al. \(2020\)](#) and [Liu et al. \(2021\)](#), the current mainstream medical report generation models use the attention mechanism to enhance the correlation between image regions and text and obtain good experimental results.

Although the medical report generation model based on the attention mechanism has become mainstream and has achieved good experimental results, it still cannot effectively solve the following problems:

1. The highly heterogeneous distribution of normal and abnormal samples in the data set (data bias) and the extremely subtle differences between X-ray images make it often difficult for data-driven neural network algorithms to generate accurate reports.
2. Medical reports usually consist of long paragraphs, and existing methods treat report generation as an isolated process, ignoring the impact of context on wording, and therefore cannot maintain contextual consistency.

In this work, we propose a new auxiliary signal guidance and memory-driven (ASGMD) network to solve the above problems. This network has up-to-date mainstream image caption architecture consisting of an encoder and decoder. In the coding part, the ASG module alleviates visual data bias by introducing auxiliary signals that are used to enhance the abnormal area features of the image. The decoding stage includes two modules, TSAM and MMDD, that maintain the consistency of the medical report text, thereby alleviating textual data bias, and fuse medical image features for dynamic decoding to optimize traditional decoders and thereby obtain more accurate and smooth medical reports.

In summary, the main contributions of this paper are as follows:

1. We propose a novel signal-guided and memory-driven network for end-to-end medical report generation.
2. To alleviate the problem of data bias, we introduce an auxiliary signal that guides the module in capturing the abnormal area within the image, yields better coding information for the abnormal area of the medical image, and enhances the image coding ability.
3. To enhance the semantic association between image and text and obtain abundant text feature representation simultaneously, we propose a novel TSAM module. In addition, we design an MMDD module that is used to dynamically generate diagnostic reports.
4. The experimental results demonstrate that our method outperforms current state-of-the-art models when used on the public IU X-ray and MIMIC-CXR datasets. The generated report is also obviously clearer and more fluent.

The remainder of this article is organized as follows. Section 2 reviews related studies on the automatic generation of medical reports. Section 3 presents the method proposed in this paper. Section 4 describes the experimental details of our experiments. Section 5 summarizes the work presented in this paper.

## 2. Related work

This section will discuss two types of work related to our work, image captioning and radiology report generation.

### 2.1. Image captioning

Image captioning aims to automatically generate corresponding descriptive text from the visual information contained in the image. Because it involves multiple research directions, including computer vision (CV) and natural language processing (NLP), it has received extensive attention from academic workers and has achieved great

success. In the early days of computer vision development, most work used retrieval or template-based methods to generate descriptions of images. [Farhadi et al. \(2010\)](#) use the nearest neighbor method to select images and Tree-F1 rules to match their corresponding descriptive texts to generate sentences. Tree-F1 reflects the accuracy and specificity of generated sentences and real sentences. [Kulkarni et al. \(2011\)](#) used conditional random fields to extract descriptive information on the word composition that best matches the image description in the database. Since retrieval and template-based methods can only generate text under a subtitle word or within a fixed template, the methods described above for generating descriptions of images are difficult to apply in practice. [Sun et al. \(2015\)](#) proposed image-based visual discrimination to filter text terms, group them using semantic and visual similarities, and then used bidirectional retrieval to generate descriptions.

Recently, inspired by machine translation tasks, encoder-decoder architecture has become the standard paradigm for image captioning tasks. In the encoder part, the convolutional neural network is usually used as the encoder of visual features. The decoder uses a recurrent neural network (RNN or LSTM) to generate descriptive information. [Vinyals et al. \(2015\)](#) first proposed the use of encoder-decoder architecture to address the image captioning problem. A CNN is used to extract the global features of the images, and long short-term memory (LSTM) networks are used to generate captions. [Xu et al. \(2015\)](#) focused on dynamic visual features by incorporating an attention mechanism and used LSTM to generate the final image description. [Krause et al. \(2017\)](#) used hierarchical LSTMs to decode visual features to generate dense image descriptions. [Huang et al. \(2019\)](#) proposed an attention mechanism that determines the correlation between attention results and queries for caption generation. [Zhou et al. \(2020\)](#) fused image and text features into a modified BERT network and utilized multiple transformer encoders for visual encoding and text generation. [Yang, Wu, et al. \(2022\)](#) used a knowledge-enhanced multi-head attention mechanism for caption generation. [Bae et al. \(2022\)](#) proposed using words corresponding to specific parts of speech to control the generation of subtitles. [Zhou et al. \(2022\)](#) proposed a collaborative strategy network integrating generative adversarial networks and spatial attention to generate descriptive information. [Wang and Gu \(2022\)](#) proposed a global contextual attention mechanism for captioning. [Jiang et al. \(2022\)](#) proposed a hybrid attention model to generate captions.

### 2.2. Medical report generation

Image captioning and medical report generation tasks take visual features as input and generate corresponding textual information. However, unlike the situation with image captioning, the regional feature differences between medical images are often slight and lack sufficient object-level supervision information. Therefore, it is often difficult to achieve effective results using an image captioning model for medical report generation. [Jing et al. \(2018\)](#) established a multitask learning framework that jointly performed label prediction and paragraph generation and proposed a coattention mechanism for generating medical reports using hierarchical LSTMs. [Li et al. \(2018\)](#) proposed a model that mixes template retrieval and normal and abnormal text generation, enhancing the model's ability to describe abnormal images. [Xue et al. \(2018\)](#) proposed a multimodal recursive model that incorporates an iterative decoder for visual attention to improve text coherence. [Li et al. \(2019\)](#) proposed using a manually extracted template database to aid in report generation; this method reconciled traditional knowledge-based and retrieval-based methods with modern learning-based methods to generate more accurate and robust medical reports. [Jing et al. \(2019\)](#) introduced two RNNs as "usual" and "exception" report writers to alleviate the data bias problem. Although this category of methods has achieved specific results, it lacks contextual consistency, and the generated paragraphs contain repeated sentences. [Li et al. \(2022\)](#) proposed an auxiliary signal-guided knowledge encoder-decoder structure

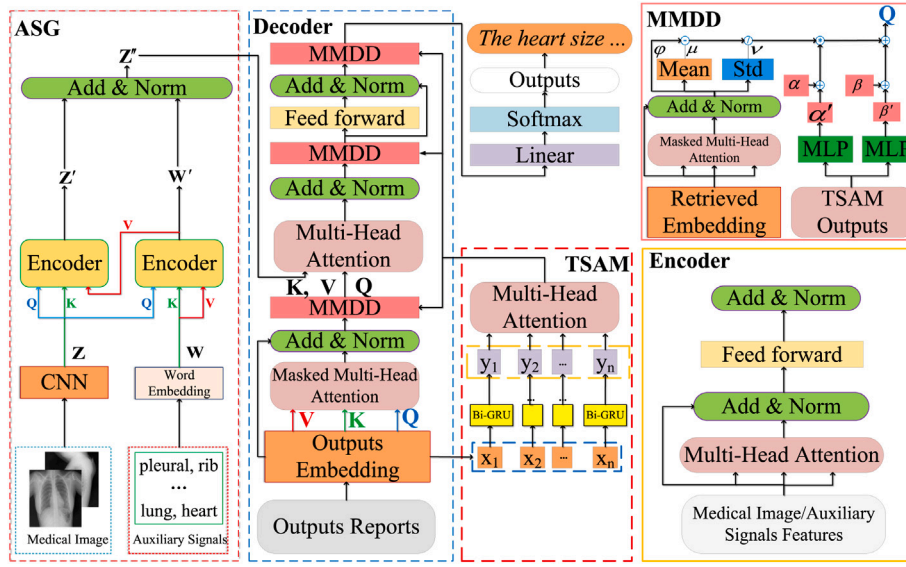


Fig. 1. Architecture of the ASGMD network.

integrating internal visual feature fusion and external medical language information, it is prone to the problems of false superposition and vanishing gradient. Recently, medical diagnosis report generation based on CNN-transformer architecture has become a new research hotspot. [Chen et al. \(2020\)](#) proposed a CNN-transformer-based medical image report generation framework and designed a storage mechanism for recording important information. [Miura et al. \(2021\)](#) proposed using two additional metrics to enhance the consistency and completeness of the report and optimized the method using a reinforcement learning strategy. [Liu et al. \(2021\)](#) constructed an application that uses prior and posterior knowledge in the CNN-Transformer architecture. [Yang, Yu, et al. \(2022\)](#) proposed three parallel networks that facilitate report generation by comparing normal and abnormal sample information. [Qin and Song \(2022\)](#) used improved reinforcement learning strategies to align visual and textual features for report generation. [Kaur and Mittal \(2022\)](#) proposed a diagnostic report generation network that fuses a co-attention mechanism with reinforcement learning strategies.

Although the above methods have produced good results, most adopt CNN-RNN architecture. When RNN encodes long text information, it is prone to the problems of false superposition and vanishing gradient. Recently, medical diagnosis report generation based on CNN-transformer architecture has become a new research hotspot. [Chen et al. \(2020\)](#) proposed a CNN-transformer-based medical image report generation framework and designed a storage mechanism for recording important information. [Miura et al. \(2021\)](#) proposed using two additional metrics to enhance the consistency and completeness of the report and optimized the method using a reinforcement learning strategy. [Liu et al. \(2021\)](#) constructed an application that uses prior and posterior knowledge in the CNN-Transformer architecture. [Yang, Yu, et al. \(2022\)](#) proposed three parallel networks that facilitate report generation by comparing normal and abnormal sample information. [Qin and Song \(2022\)](#) used improved reinforcement learning strategies to align visual and textual features for report generation. [Kaur and Mittal \(2022\)](#) proposed a diagnostic report generation network that fuses a co-attention mechanism with reinforcement learning strategies.

Although the above work has achieved great success, there are still shortcomings in addressing visual data bias and in the consistency of the generated reports. In this paper, we focus on solving these two problems. Specifically, we introduce the use of medical disease subject entities as auxiliary signals to enhance abnormal regions of medical images to alleviate the problem of visual data bias. Second, we build a text encoding module with a memory function to ensure the consistency of the generated reports. Finally, we integrate the text encoding information into the transformer decoder to generate the final report.

### 3. Our proposed method

#### 3.1. Overview

The essence of automatic generation of medical imaging reports is the task of mapping images to text. Following the latest standards for image captioning, we regard the image-to-text task as a sequence-to-sequence task. For a given image, we define the image features as the source sequence  $\mathbf{Z} = \{z_1, z_2, z_3, \dots, z_K\}$ ,  $z_K \in \mathbb{R}^d$ , where  $z_K$  are patch features extracted from visual extractors and  $d$  represents the dimension of the feature vector. The text sequence corresponding to

the image is  $\mathbf{T} = \{t_1, t_2, t_3, \dots, t_N\}$ ,  $t_N \in \mathbb{V}$ , where  $t_N$  are the generated text sequences and  $N$  and  $\mathbb{V}$  represent the sequence length and the set of all possible words, respectively.

An overview of our proposed ASGMD model is shown in [Fig. 1](#). The model includes three main modules: auxiliary signal guidance (ASG), text sequential attention mechanism (TSAM), and memory mechanism-driven decoding (MMDD).

When radiologists examine medical images of patients, they usually focus on the abnormal areas in the images, then quickly browse the global information and write corresponding reports. To mimic the above working pattern, we selected the top 20 disease subject entities (i.e. strong medically relevant signal) according to the frequency of word occurrences in the training set report as auxiliary signals of the ASG module to enhance image features and thereby alleviate the visual data bias problem. It is worth mentioning that the auxiliary signal part of our design was inspired by the work of [Liu et al. \(2021\)](#), which designed a more complex knowledge graph and selected auxiliary information from it. However, building knowledge graphs is often difficult and not general. Also, the experimental results prove that our method is better than the results of [Liu et al. \(2021\)](#). The auxiliary signal is defined as  $\mathbf{W} = \{\text{normal, cardiomegaly, scoliosis, fractures, effusion, thickening, pneumothorax, hernia, calcinosis, emphysema, pneumonia, edema, atelectasis, catatrix, opacity, lesion, airspace disease, hypoinflation, medical device, other}\}$ . Second, diagnostic reports usually consist of long paragraphs and sentences. Existing methods still have shortcomings in handling long texts with consistency. We propose a TSAM module that ensures the fluency and contextual consistency of the generated medical reports. Finally, the MMDD is devoted to distilling helpful information to generate reports. The MMDD module was partially inspired by [Chen et al. \(2020\)](#). In our work, to incorporate the auxiliary signals into the transformer's decoder, we refer to the memory-driven conditional layer designed by [Chen et al. \(2020\)](#). In the experimental section, we make a detailed comparison with [Chen et al. \(2020\)](#). The experimental results demonstrate that our method outperforms the method proposed by [Chen et al. \(2020\)](#).

In brief, the proposed ASGMD takes  $\mathbf{Z}$ ,  $\mathbf{W}$  as input and generates the robust report  $\mathbf{T}$ . This process can be modeled as:

$$p(\mathbf{T}|\mathbf{Z}, \mathbf{W}) = \prod_{n=1}^N p(t_n|t_1, t_2, \dots, t_{n-1}, \mathbf{Z}, \mathbf{W}) \quad (1)$$

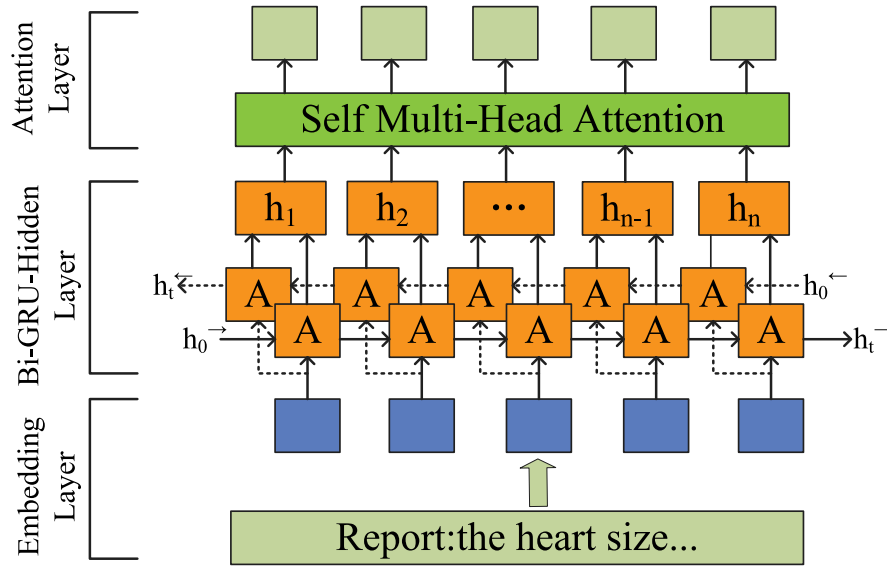


Fig. 2. The architecture of the TSAM includes three layers: the embedding layer, the Bi-GRU-hidden layer, and the attention layer.

The model is then trained to maximize  $p(\mathbf{T}|\mathbf{Z}, \mathbf{W})$  through the negative conditional log-likelihood of  $\mathbf{T}$  given  $\mathbf{Z}$  and  $\mathbf{W}$ .

$$\theta^* = \arg \max_{\theta} \sum_{n=1}^N \log p(t_n | t_1, t_2, t_{n-1}, \mathbf{Z}, \mathbf{W}; \theta) \quad (2)$$

where  $\theta$  is a parameter of the model. Given examples in the training set, we use a cross-entropy loss function to maximize the conditional log-likelihood:

$$\mathcal{L}_{CE} = \sum_{n=1}^N \log p_{\theta}(\mathbf{T}^{(n)} | \mathbf{Z}^{(n)}, \mathbf{W}^{(n)}) \quad (3)$$

### 3.2. Auxiliary signal guidance module

We use CNN as the visual feature extractor and vectorize the auxiliary signal with random initialization. The results are expressed as  $\mathbf{Z} = \{z_1, z_2, z_3, \dots, z_K\}$  and  $\mathbf{W} = \{w_1, w_2, \dots, w_{20}\}$ , where  $w_j$  refers to the word embedding of the  $j^{th}$  signal. The ASG module mainly utilizes the auxiliary signal  $\mathbf{W}$  to facilitate the representation of the original visual feature  $\mathbf{Z}$  and integrates it into the decoder as a key and value matrix to facilitate report generation. We implement the proposed method in a basic transformer model consisting mainly of multihead attention and a feed-forward network (FFN). The multihead attention consists of parallel heads, each head defined as a proportional dot product:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (4)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (5)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)\mathbf{W}^O \quad (6)$$

where  $\mathbf{Q}$ ,  $\mathbf{K}$  and  $\mathbf{V}$  represent the query, key and value matrices, respectively, and  $\mathbf{W}_i^Q \in d_{model} \times d_k$ ,  $\mathbf{W}_i^K \in d_{model} \times d_k$ ,  $\mathbf{W}_i^V \in d_{model} \times d_v$  and  $\mathbf{W}^O \in d_v \times d_{model}$  are learnable parameters. In this work, we use 8 headers. For each head, we use  $d_k = d_v = d_{model}/h = 64$  dimensions.

To add nonlinearity to the model, an FFN is applied after the attention to obtain nonlinear features:

$$\text{FFN}(\mathbf{X}) = \text{Max}(0, \mathbf{X}\mathbf{W}_f + \mathbf{b}_f)\mathbf{W}_{ff} + \mathbf{b}_{ff} \quad (7)$$

In this equation,  $\text{Max}(0, *)$  represents the ReLU activation function,  $\mathbf{W}_f \in \mathbb{R}^{d \times 4d}$  and  $\mathbf{W}_{ff} \in \mathbb{R}^{d \times 4d}$  are learnable matrices, and  $\mathbf{b}_f$  and  $\mathbf{b}_{ff}$

represent the bias term. It is worth noting that both the MHA and the FFN are followed by an operational sequence of dropout, residual connection, and layer normalization.

To this end, the ASG module is modeled as follows:

$$\mathbf{W}' = \text{FFN}(\text{MultiHead}(\mathbf{Z}, \mathbf{W}, \mathbf{W})) \quad (8)$$

$$\mathbf{Z}' = \text{FFN}(\text{MultiHead}(\mathbf{Z}, \mathbf{Z}, \mathbf{W}')) \quad (9)$$

where  $\mathbf{Z}$  and  $\mathbf{W}$  are denoted as image features and auxiliary signals, respectively.

Since the auxiliary signal  $\mathbf{W}$  contains the abnormal medical-related topics, we can not only obtain the abnormal area from the auxiliary signal but also align the abnormal area of interest with the relevant auxiliary signal (see Fig. 1).

Finally, we simply add and normalize  $\mathbf{Z}'$  and  $\mathbf{W}'$  to represent the final visual feature:

$$\mathbf{Z}'' = \text{layernorm}(\mathbf{Z}' + \mathbf{W}') \quad (10)$$

where layernorm is denoted as layer normalization (Ba et al., 2016).

### 3.3. Text sequential attention mechanism

Medical diagnostic reports usually consist of long paragraphs that describe the patient's condition, and it is still a great challenge to maintain the consistency of coding context in long texts. To correctly express medical knowledge, we propose using an auxiliary component, the TSAM module, to optimize the diagnostic report generation process. The TSAM module is responsible for extracting global information from the medical report created from the training data, as shown in Fig. 2. Given the excellent performance of bidirectional GRU (Cho et al., 2014) in text encoding work, we use it as a precursor encoding method for diagnostic reports. The first layer of TSAM is the embedding layer; it maps each word of the diagnostic report to a dimensional vector and then converts the vector into a fixed-length embedding matrix. The second layer is the Bi-GRU-Hidden layer. The number of neurons in the hidden layer is 768, and the output for each time step is defined as follows:

$$z_t = \text{sigmoid}(h_{t-1}\mathbf{W}_z + \mathbf{X}_t\mathbf{U}_z) \quad (11)$$

$$r_t = \text{sigmoid}(h_{t-1}\mathbf{W}_r + \mathbf{X}_t\mathbf{U}_r) \quad (12)$$

$$\tilde{h}_t = \tanh(\mathbf{W}\mathbf{X}_t + \mathbf{U}(h_{t-1} \odot r_t)) \quad (13)$$



$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \quad (14)$$

where  $h_t$  and  $x_t$  are the hidden layer states at moment  $t$ ,  $h_{t-1}$  is the hidden state of each layer at moment  $t-1$  or the initial hidden state at moment 0.  $z_t$ ,  $r_t$ , and  $\tilde{h}_t$  represent the update, the reset, and the new gate, respectively.  $W_*$  and  $U_*$  are the learnable weight matrices,  $\odot$  is the Hadamard product.

We concatenate the output correspondence of the bidirectional hidden layer as the final result:

$$\mathbf{H} = (h_1, h_2, \dots, h_n) \quad (15)$$

where  $h_i$  represents the encoding vector of the  $i^{th}$  word.

Although bidirectional GRU delivers excellent performance in encoding text, it still has nontrivial limitations in solving the problem of long text invalidation. Therefore, we propose the use of multiheaded self-attention to solve this problem. It is worth mentioning that to the best of our knowledge this is the first study in which bidirectional GRU + multihead self-attention is used to solve the long text encoding problem and achieve substantial results. The equation is as follows:

$$\mathbf{M} = \text{Muitlhead}(\mathbf{H}, \mathbf{H}, \mathbf{H}) \quad (16)$$

where  $\mathbf{H}$  is the result of concatenating all the hidden states of the bidirectional GRU.

### 3.4. Memory mechanism-driven decoding module

Although the memory module has succeeded in NLP tasks, as shown by Sukhbaatar et al. (2015) and Kumar et al. (2016), most existing NLP tasks apply it to designing independent codes.

Given that medical report generation is a dynamic process that is affected by the output of each decoding step, we propose the use of an MMDD module to fuse the memory information  $\mathbf{M}$  and retrieve the report generated by the TSAM module (see Fig. 1). The MMDD module consists of two branches. The left branch is used to calculate the variance  $v$  and the standard deviation  $\mu$  of the original text features; the right branch uses the  $v$  and  $\mu$  values calculated by the left branch to change the two simple parameters  $\alpha$  and  $\beta$  in the traditional transformer. In this way, the contextual consistency of the generated report is maintained without excessive loss of original text feature information. As shown in Fig. 1, we use three MMDD modules in each transformer decoding layer; the output of the first MMDD is functionalized as the query to be fed into the following multihead attention module together with the hidden states from the encoder as the key and value. Specifically, we feed the memory information  $\mathbf{M}$  into the MMDD module, and an MLP is used to predict the change of  $\alpha$  to  $\alpha'$  and the change of  $\beta$  to  $\beta'$ . The updates are as follows:

$$\alpha' = F_{mlp}(\mathbf{M}), \tau_\alpha = \alpha + \alpha' \quad (17)$$

$$\beta' = F_{mlp}(\mathbf{M}), \tau_\beta = \beta + \beta' \quad (18)$$

Finally,  $\tau_\alpha$  and  $\tau_\beta$  are applied to the mean and variance results of the multihead self-attention from the previously generated outputs:

$$F_{MMDD}(\varphi) = \tau_\alpha \odot \frac{\varphi - \mu}{v} + \tau_\beta \quad (19)$$

In this equation,  $\varphi$  refers to the output of the previous module, and  $\mu$  and  $v$  represent the mean and the standard deviation, respectively, of  $\varphi$ .

## 4. Experiments

### 4.1. Datasets

We conduct experiments on two public datasets, MIMIC-CXR (Johnson et al., 2019) and IU X-ray (Demner-Fushman et al., 2016).

**Table 1**

Data statistics for the MIMIC-CXR and IU X-ray.

Dataset	Type	Train	Val	Test
IU X-ray	Images	5,226	748	1,496
	Reports	2,770	395	790
	Average length	37.56	36.78	33.62
MIMIC-CXR	Image	368,960	2,991	5,159
	Report	222,758	1,808	3,269
	Average length	53	53.05	66.4

MIMIC-CXR is the largest radiology dataset available, comprising 473,057 chest X-ray images and 206,563 reports.

IU X-ray is a widely used benchmark dataset. It contains 7,470 chest X-ray images associated with 3,955 radiology reports.

The data statistics are shown in Table 1. The statistics include the number of images, the number of reports, and the average length of the reports. We first exclude unreported samples. IU X-ray is then divided into training, validation, and test sets in a ratio of 7:1:2 for the entire dataset. For the MIMIC-CXR dataset, we adopt the official division for a fair comparison. Specifically, the MIMIC-CXR dataset is split into 368,960 images/222,758 reports for training, 2,991 images/1,808 reports for validation and 5,159 images/3,269 reports for testing. For the MIMIC-CXR dataset, the maximum word length is 100, which contains 7870 words. In addition, the IU X-ray dataset's maximum word length is 60, which includes 767 words. Our model consists mainly of a transformer, Bi-GRU, and convolutional neural network with 118,911,497 (453MB) trainable integer parameters.

### 4.2. Parameter setting and evaluation metrics

The experiments in this paper were implemented on an Intel Xeon Gold 5218 CPU @2.3 GHz, 128 GB RAM, and NVIDIA Tesla V100 SXM2 GPU. We set ResNet-101 (He et al., 2016) and Resnet152 (He et al., 2016) as image feature extractors for the IU X-ray and MIMIC-CXR datasets. The obtained image feature  $Z \in \mathbb{R}^{2048 \times 7 \times 7}$  is mapped to  $Z \in \mathbb{R}^{512 \times 49}$ . Notably, for the IU X-ray dataset, we feed the front/side images of the patient into the convolutional neural network in parallel and fuse them into appropriate image features. In the MMDD module,  $\alpha'$  and  $\beta'$  are received by two MLPs. We use the leveraging ADAM optimizer to perform label classification on the model under cross-entropy loss during training. We used the grid-search method to find the optimal parameters within the given parameter range. For image feature extractors on IU X-ray and MIMIC-CXR data, we set the learning rate in the range [5e-1, 5e-2, 5e-3, ..., 5e-9], and the optimal parameters are 5e-5 and 5e-2, respectively. The learning rate of the model ranges from [1e-1, 1e-2, 1e-3, ..., 1e-6], and the optimal model learning rate is 1e-5. The optimal solutions for dropout are 0.1 and 0.3, respectively. The optimal regularization parameters are 0.0001 and 0.0008, respectively.

**Metrics:** We adopt three standard metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004), all of which are calculated by the standard evaluation toolkit. BLEU is used to evaluate the similarity between sentence pairs. The specific formula is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (20)$$

$$w_n = \frac{1}{n} \quad (21)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (22)$$

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{n\_gram \in C} count_{clip}(n\_gram)}{\sum_{C' \in \{candidates\}} \sum_{n\_gram' \in C'} count(n\_gram')} \quad (23)$$

$$count_{clip} = \min(count, max\_ref\_count) \quad (24)$$

**Table 2**

Comparison of the MIMIC-CXR and IU X-ray datasets. Specifically, we quote the results reported by Liu et al. (2021) for MIMIC-CXR and IU X-ray. The results shown in bold type indicate the best performance. The P values denote the significance levels of the differences according to the T-test. The results of our model are averaged from five experiments.

Dataset	Model	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MIMIC-CXR	CNN-RNN (Vinyals et al., 2015)	0.124	0.263	0.299	0.184	0.121	0.084
	Att2in (Rennie et al., 2017)	0.134	0.276	0.325	0.203	0.136	0.096
	TOP-Down (Anderson et al., 2018)	0.128	0.267	0.317	0.195	0.13	0.092
	Transformer (Chen et al., 2020)	0.125	0.265	0.314	0.192	0.127	0.09
	R2Gen (Chen et al., 2020)	0.142	0.277	0.353	0.218	0.145	0.103
	PPKED (Liu et al., 2021)	0.149	0.284	0.36	0.224	0.149	0.106
	CMN (Chen et al., 2021)	0.142	0.278	0.353	0.218	0.148	0.106
	Ours	<b>0.152</b>	<b>0.286</b>	<b>0.372</b>	<b>0.233</b>	<b>0.154</b>	<b>0.112</b>
	P value	0.00068891	0.0031982	2.88E-06	0.00014162	9.35E-05	0.0031982
IU X-ray	CoAtt (Jing et al., 2018)	-	0.369	0.455	0.288	0.205	0.154
	SentSAT+KG (Zhang et al., 2020)	-	0.367	0.441	0.291	0.203	0.147
	Transformer (Chen et al., 2020)	0.164	0.342	0.396	0.254	0.179	0.135
	R2Gen (Chen et al., 2020)	0.187	0.371	0.47	0.304	0.219	0.165
	PPKED (Liu et al., 2021)	0.19	0.376	0.483	0.315	0.224	0.168
	CMN (Chen et al., 2021)	0.191	0.375	0.475	0.309	0.222	0.17
	Ours	<b>0.206</b>	<b>0.397</b>	<b>0.489</b>	<b>0.326</b>	<b>0.232</b>	<b>0.173</b>
	P value	1.18E-06	3.08E-07	0.00149353	4.08E-06	0.11669147	0.0132356

where  $w_n$  represents the weight of  $n\_gram$ ,  $n \in \{1, 2, 3, 4\}$ ,  $r$  and  $c$  are the lengths of the reference translation and the candidate translation, respectively,  $count_{clip}(n\_gram)$  represents the number of occurrences of  $n\_gram$  in the reference,  $count(n\_gram')$  represents the number of  $n\_gram'$  occurrences in the candidate.  $count$  represents the number of occurrences of  $n\_gram$  in the candidate, and  $max\_ref\_count$  is the maximum number of occurrences of  $n\_gram$  in the reference.

METEOR is a recall metric for machine translation. The use of this recall metric overcomes the disadvantage that BLEU cannot locate semantic similarity. The specific formula is as follows:

$$METEOR = (1 - pen) \times F_{means} \quad (25)$$

$$pen = \frac{\#chunks}{m} \quad (26)$$

$$F_{means} = \frac{PR}{\alpha P + (1 - \alpha)R} \quad (27)$$

$$P = \frac{m}{c} \quad (28)$$

$$R = \frac{m}{r} \quad (29)$$

where  $pen$  is the penalty factor, and  $\#chunks$  represents the number of chunks (the unit of aggregation of matched 1-tuples that are adjacent in both the candidate and the reference translations).  $\alpha$  is a controllable parameter,  $m$  is the number of tuples that can be matched in the candidate translation,  $c$  is the length of the candidate translation, and  $r$  is the length of the reference translation.

ROUGE-L is responsible for computing the longest common subsequence of sentences. The specific formula is as follows:

$$ROUGE - L = \frac{(1 + \beta^2)R_{lcs}P_{lsc}}{R_{lcs} + \beta^2 P_{lsc}} \quad (30)$$

$$R_{lcs} = \frac{LSC(X, Y)}{m} \quad (31)$$

$$P_{lsc} = \frac{LSC(X, Y)}{n} \quad (32)$$

In these equations,  $X$  represents the candidate translation,  $Y$  represents the reference translation,  $LSC(X, Y)$  is the length of the longest common subsequence of the candidate translation and the reference translation, and  $m$  and  $n$  represent the length of the reference translation and the length of the candidate translation, respectively.

#### 4.3. Quantitative results

We compare the method proposed in this paper with previous state-of-the-art models, i.e., CNN-RNN (Vinyals et al., 2015), Att2in (Rennie et al., 2017), CoAtt (Jing et al., 2018), TOP-Down (Anderson et al., 2018), SentSAT+KG (Zhang et al., 2020), Transformer (Chen et al., 2020), R2Gen (Chen et al., 2020), PPKED (Liu et al., 2021), and CMN (Chen et al., 2021). It is worth noting that, to ensure the fairness of the comparison, the results obtained for our comparison objects are all obtained by comparison with the results of the original papers, as shown in Table 2.

The details of the above models are as follows:

1. CNN-RNN was the first CNN-RNN architecture for solving image captioning. CNN acts as a visual feature extractor, and RNN acts as a decoder to generate textual information.
2. Att2in performs captioning generation via a self-critical sequence training (SCST) reinforcement learning strategy.
3. In CoAtt, a multitask learning framework is proposed. A collaborative attention mechanism is used to fuse visual features with semantic features of medical labels, and a hierarchical LSTM is used to generate medical diagnosis reports.
4. In TOP-Down, a bottom-up and top-down attention mechanism is proposed. Bottom-up (based on Faster R-CNN) is used to extract regional features of the image, and top-down is used to calculate the weights of relevant feature vectors and to finally encode the features to generate textual information.
5. In SentSAT+KG, medical report generation is performed by pre-constructing a knowledge graph that is based on a graph convolutional neural network related to medical topics.
6. Transformer generates medical reports based on the traditional transformer framework. CNNs are used to extract visual features, and those features are decoded using standard transformer decoders.
7. R2Gen uses CNN-Transformer architecture to generate medical reports. Unlike the transformer, R2Gen includes an additional component (RM) that records the generated information and fuses the information in the RM into the transformer decoder to facilitate report generation.
8. In PPKED, visual features are first extracted by CNN, and the constructed prior and posterior knowledge are then integrated into a standard transformer to facilitate report generation.
9. CMN proposes a cross-modal memory network (CMN) for medical report generation. CNN is used to extract visual features,





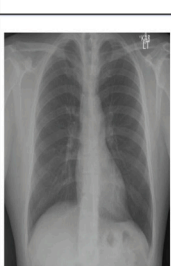
Original Image	Ground Truth	R2Gen	Ours
	normal cardiomeastinal silhouette . there is no focal consolidation . there are no xxxx of a large pleural effusion . there is no pneumothorax . there is no acute bony abnormality seen .	the heart size and mediastinal silhouette are within normal limits for contour . there are no xxxx of a large pleural effusion . the xxxx are intact .	the heart size and pulmonary vascularity appear within normal limits . there is no focal consolidation .there is no pneumothorax . there is no acute bony abnormality seen .
	lungs are clear . there is no pneumothorax or pleural effusion . the heart and mediastinum are within normal limits . bony structures are intact .	mediastinal contours are within normal limits . heart size is within normal limits . no focal consolidation pneumothorax or pleural effusion .	lungs are clear . no pleural effusions or pneumothoraces . heart and mediastinum of normal size and contour . degenerative changes in the spine .
	cardiac and mediastinal contours are within normal limits . the lungs are clear . bony structures are intact .	lungs are clear . no pleural effusions or pneumothoraces . heart and mediastinum of normal size and contour . degenerative changes in the spine .	lungs are clear . no pneumothorax or pleural effusion . cardiac and mediastinal contours are within normal limits .
	The heart is normal size with normal appearance the cardiomeastinal silhouette. There is no focal air space opacity, pleural effusion, or pneumothorax. The osseous structures are intact with degenerative changes in thoracic spine.	the heart is normal size . no pleural effusions or pneumothoraces .	the heart size and cardiomeastinal silhouette are normal . the lungs are clear without focal airspace or pneumothorax .
	the cardiomeastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Osseous structures are within normal limits for patient age.	lungs are clear without focal consolidation effusion or pneumothorax. normal heart size .	the cardiomeastinal silhouette is within normal limits. heart and mediastinum normal. no pneumonia effusions edema pneumothorax adenopathy nodules or masses .

Fig. 3. Comparison of reports generated by our method and by the R2Gen model of Chen et al. (2020) The leftmost column displays the ground truth; the middle column shows the medical diagnosis report generated by the R2Gen method, and the column on the right shows the medical diagnosis report generated by our method. Different colors are used to distinguish the reports generated by the R2Gen method and by our method and the ground truth.

and an additional storage module is designed to record the semantic alignment information of visual and textual features. This information is finally decoded by a standard transformer decoder.

10. Base: Our baseline model follows the current standard paradigm in this task, i.e., it uses a CNN as a visual feature extractor and a standard transformer for report generation.

Our method outperforms state-of-the-art methods on the MIMIC-CXR and IU X-ray datasets. In particular, on the IU X-ray dataset, all evaluation metrics have show an average improvement of 4.18%. The performance of the ROUGE-L evaluation metric has improved by 5.6%. Notably, our method achieves an 8.4% performance improvement on the METEOR evaluation metric compared to the state-of-the-art model. On the MIMIC-CXR dataset, the overall performance is improved by an average of 3.15%, and the performance according toon the BELU-4 evaluation metric is improved by 5.6%. This shows that our method

effectively alleviates the data bias problem and that it can capture abnormal regions in medical images and generates accurate diagnostic reports. In addition, to test the significance of our experimental results, we use T-test to compute the P value of our model results and the mean of the optimal baseline (Liu et al., 2021) results. As shown in the Table 2, the P value is less than 0.05, proving our model's significance.

#### 4.4. Qualitative results

To further verify the quality of the generated reports, we analyze the reports generated by the ASGMD model in this paper and the R2Gen model, as shown in Fig. 3. The ground truth represents the accurate report on the image. We use blue and red to describe the consistency between the actual report and the report generated by R2Gen (Chen et al., 2020) and the ASGMD models, respectively. Compared with R2Gen, the report generated by our method has a higher degree of

**Table 3**  
Ablation studies of the ASGMD network.

Dataset	Model	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MIMIC-CXR	Base	0.088	0.248	0.166	0.104	0.071	0.05
	Base+ASG	0.108	0.256	0.261	0.163	0.107	0.074
	Base+TSAM	0.184	0.251	0.267	0.167	0.116	0.085
	Base+TSAM+MMDD	0.137	0.265	0.331	0.213	0.144	0.102
	<b>Ours</b>	<b>0.152</b>	<b>0.286</b>	<b>0.372</b>	<b>0.233</b>	<b>0.154</b>	<b>0.112</b>
IU X-ray	Base	0.158	0.336	0.378	0.225	0.167	0.127
	Base+ASG	0.195	0.371	0.482	0.308	0.213	0.162
	Base+TSAM	0.195	0.346	0.346	0.209	0.15	0.114
	Base+TSAM+MMDD	0.188	0.343	0.309	0.203	0.142	0.104
	<b>Ours</b>	<b>0.206</b>	<b>0.396</b>	<b>0.489</b>	<b>0.326</b>	<b>0.232</b>	<b>0.173</b>

coincidence with the ground truth and is closer to people's reading and writing habits. In summary, our reports conform to the habits of radiologists in writing medical diagnoses and have good accuracy and fluency.

#### 4.5. Ablation study

We conducted ablation experiments on two datasets to verify the effectiveness of the critical modules proposed in this paper. The results are shown in Table 3.

In the ablation study of the ASG module, we first analyze the effectiveness of the module. In both datasets, each indicator yields a certain degree of improvement compared with the introduction of the ASG module in the base. The visualization result of ASGMD on the IU X-ray dataset is shown in Fig. 4. The generated reports demonstrate significant alignment with the ground truth reports as well as correspondence with the visualized attention maps. The image regions that the model focuses on are often random and meaningless until the auxiliary signals are added. By adding auxiliary signals, the model is able to focus significantly on the important regions of the image (the degree of redness indicates the degree of attention). The results show that ASGMD can generate accurate reports of interpretable attention regions.

An ablation study in which the TSAM module was ablated was performed. Table 3 shows the effectiveness of the TSAM module in our method. When the TSAM module was used on the MIMIC-CXR data, every metric improved significantly, especially BLEU-1, which improved by 10.1%. When the TSAM module was used on the IU X-ray data, both the METEOR and the ROUGE-L metrics improved.

To verify the effectiveness of the bidirectional GRU + multihead attention proposed in this paper, we set up GRU, bidirectional GRU (BiGRU), multihead attention (MHA), and their combinations, as shown in Table 4. The results of these experiments show that our proposed bidirectional GRU+multihead attention captures the semantic associations of medical reports and decreases the semantic gap between image texts. In addition, the performance of bidirectional GRU+multihead attention far exceeds that of GRU and GRU+multihead attention.

In an ablation study of the MMDD module, we explored the effectiveness of fusion of the TSAM and MMDD modules, as shown in Table 3. Fusion of the two modules achieves excellent experimental results. With the stacking of the modules, the model's overall performance in the MIMIC-CXR dataset gradually improves. The results show that the three main modules proposed in this paper all contribute to the task of generating medical reports. For BLEU-1, BLEU-2, and BLEU-4, compared with the base, the improvement is 124%, and the METEOR evaluation metric is also greatly improved. In the IU X-ray dataset, the BLEU metric is greatly affected by the ASG module. The TSAM module and its combination with the MMDD module significantly affect the METEOR and ROUGE-L metrics.

#### 4.6. Parameters study

This section describes the results of sensitivity experiments that were performed to explore the effects of parameters such as learning rate and auxiliary signal on the performance of the model.

In the parameter analysis of the learning rate, we set different learning rates for models to verify their effects on evaluation metrics, as shown in Fig. 5. On the MIMIC-CXR dataset, our method works best when the learning rate is  $5e-2$ . In addition, our approach has good adaptability to parameters, i.e., the evaluation index is less affected by parameter changes, and the performance is relatively stable. Our method achieves the best results on the IU X-ray dataset when the learning rate is  $5e-5$ . Although the method proposed in this paper achieves different results when different parameters are chosen, the overall performance of our method is better than that of the currently popular current medical report generation algorithms (see Fig. 6).

For parameter analysis of the auxiliary signal, we used six groups of auxiliary signals to verify the influence of auxiliary signals and their lengths on the model, as shown in Fig. 7. Signal 1 and signal 2, which have lengths of 20 and 30, respectively, are the medical-related auxiliary signals we selected (the first 20 signals are the same). Signal 3 and Signal 4 are weak medical-related auxiliary signals. Signals 5 and 6 are strong medically relevant auxiliary signals in the experimental setup of our model.

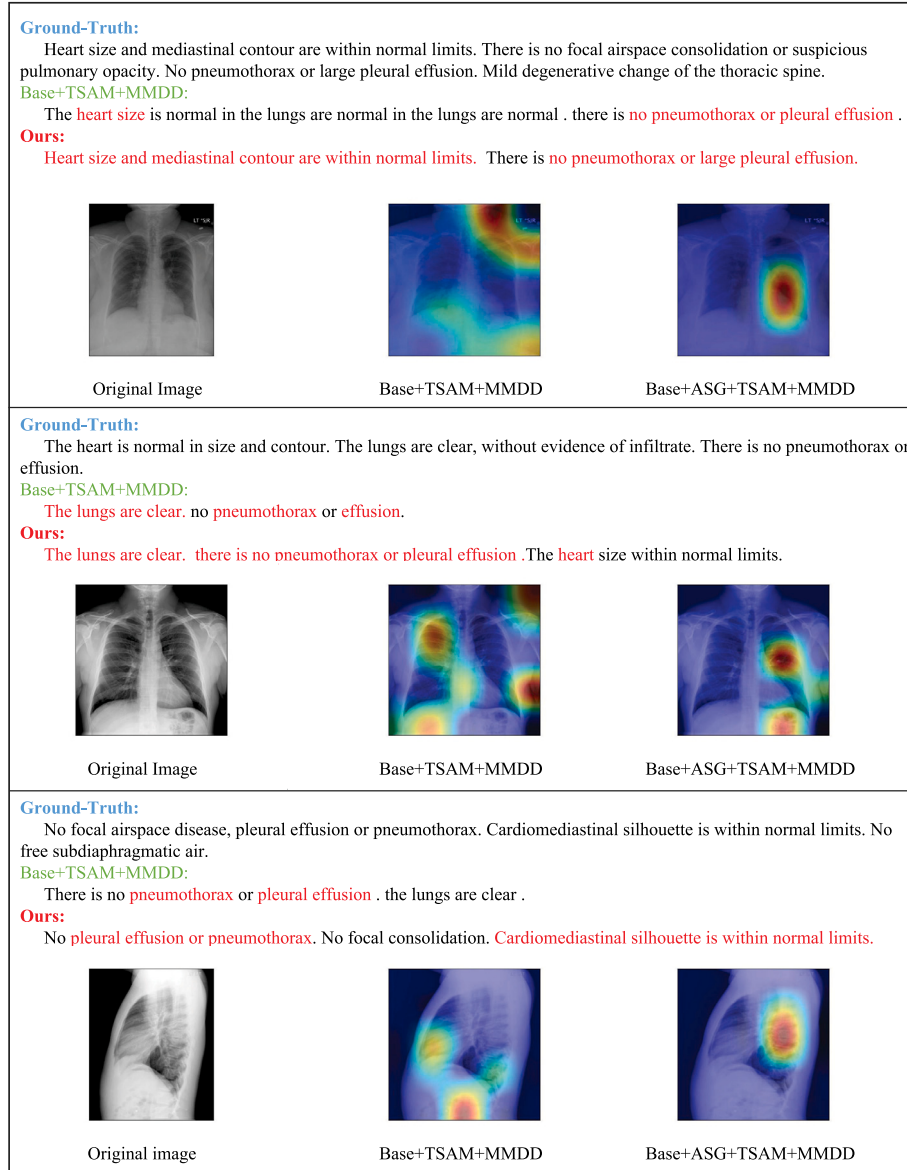
It was found that the auxiliary signals have an impact on the model. For example, the impact of strong medical-related signals is better than that of weak medical-related signals, and the experimental results of medical-related solid signals are better than those of medical-related auxiliary signals. Fig. 7 also shows that changes in the number of auxiliary signals do not help or harm the generation of sentences. For example, signal 1 and signal 2 generate precisely the same experimental results.

As shown in the Table 5, we further analyze the effect of the number of attention heads on the ASGMD model. We set the number of attention heads to 2, 4, 6, 8, and 10, respectively. It can be seen from the experimental results that different numbers of attention heads will have different effects on the evaluation metrics of the model. However, when the number of attention heads is set to 8, the model outperforms other head numbers. Therefore, we set the number of attention heads of the model to 8.

## 5. Conclusion and feature work

We propose an auxiliary signal guidance and memory-driven network called ASGMD for automatic generation of medical reports. This network aims to improve the accuracy and logic of medical reports. The ASG module is designed to capture abnormal areas in medical images and alleviate the problem of data bias. At the same time, the correlation between long texts is enhanced by development of the TSAM module, and this is beneficial for subsequent text generation. Moreover, we introduce the MMDD module as a link to the TSAM to achieve better interaction of the encoding module and the decoding module.





**Fig. 4.** Visualization of the ASGMD results in the IU X-ray dataset. Red text indicates alignment between the generated text and the ground truth report. Base + ASG + TSAM + MMDD represents our complete model; Base + TSAM + MMDD means that the auxiliary signal is not integrated. It can be clearly seen that the report generated after integration of the auxiliary signal is more fluent and accurate.

**Table 4**

Impact of different text encodings on overall performance with the TSAM.

Dataset	Methods	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MIMIC-CXR	GRU	0.062	0.182	0.296	0.167	0.102	0.074
	BiGRU	0.076	0.215	0.307	0.168	0.121	0.081
	GRU+MHA	0.122	0.249	0.351	0.211	0.129	0.101
	BiGRU+MHA	0.152	0.286	0.372	0.233	0.154	0.112
IU X-ray	GRU	0.128	0.264	0.436	0.206	0.169	0.138
	BiGRU	0.144	0.341	0.458	0.248	0.172	0.144
	GRU+MHA	0.182	0.352	0.462	0.289	0.199	0.165
	BiGRU+MHA	0.206	0.396	0.489	0.326	0.232	0.173

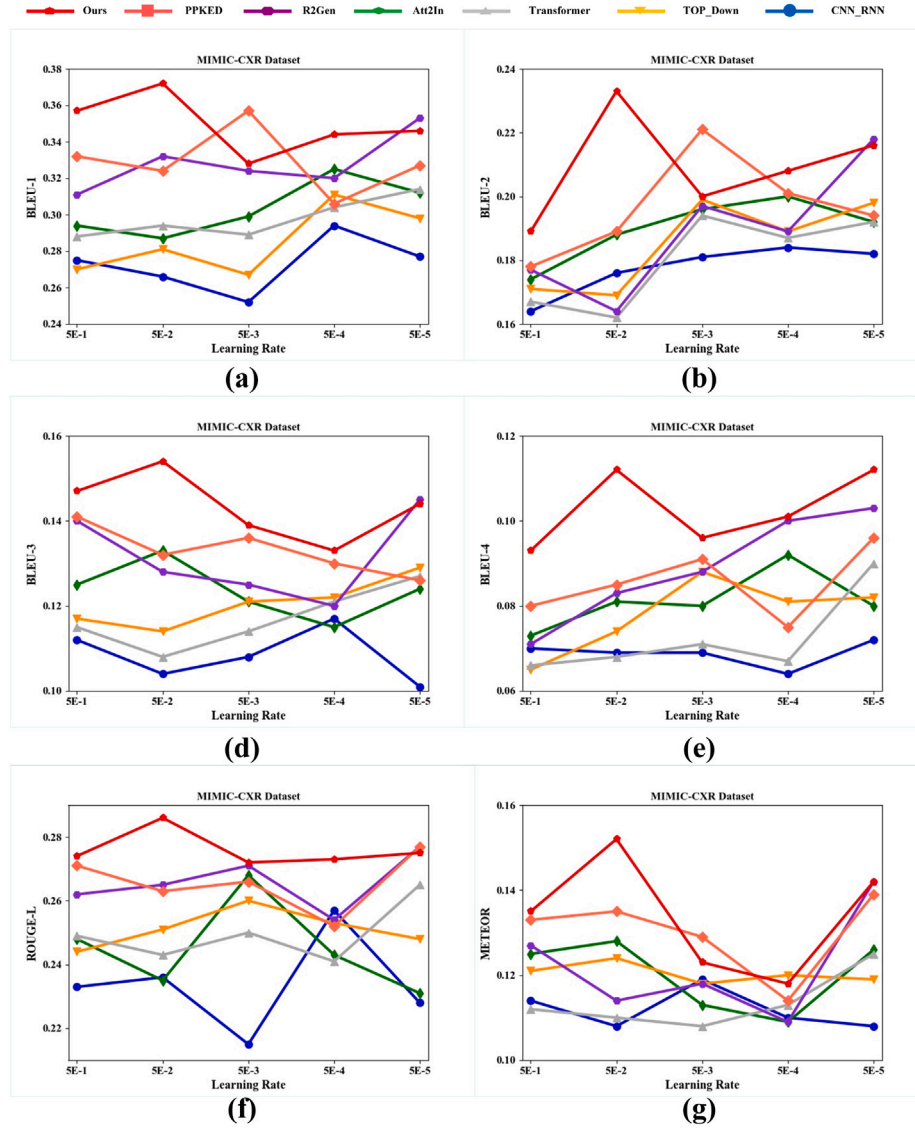


Fig. 5. Parameter sensitivity of the model on the MIMIC-CXR dataset.

Table 5

Impact of different attention heads on the overall performance of ASGMD. The results shown in bold type indicate the best performance.

Dataset	HEAD	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MIMIC-CXR	2-heads	0.147	0.279	0.365	0.212	0.142	0.098
	4-heads	0.150	0.280	0.370	0.221	0.148	0.103
	6-heads	0.148	0.283	0.369	<b>0.233</b>	0.151	0.111
	8-heads	<b>0.152</b>	0.284	<b>0.372</b>	<b>0.233</b>	<b>0.154</b>	<b>0.112</b>
	10-heads	0.151	<b>0.286</b>	0.371	0.229	0.144	0.108
IU X-ray	2-heads	0.199	0.368	0.472	0.301	0.221	0.162
	4-heads	0.197	0.369	0.481	0.321	0.234	0.166
	6-heads	0.201	0.377	0.482	0.322	0.231	0.164
	8-heads	0.206	<b>0.397</b>	<b>0.489</b>	0.326	<b>0.232</b>	<b>0.173</b>
	10-heads	<b>0.208</b>	0.371	0.487	<b>0.328</b>	0.220	0.169

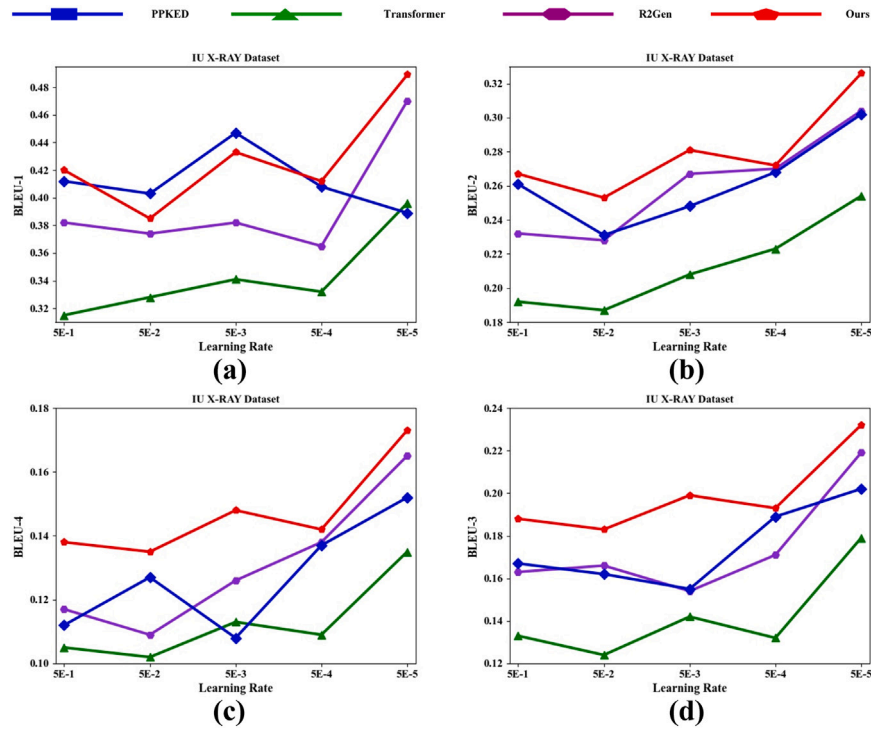


Fig. 6. Parameter sensitivity of the model on the IU X-ray dataset.

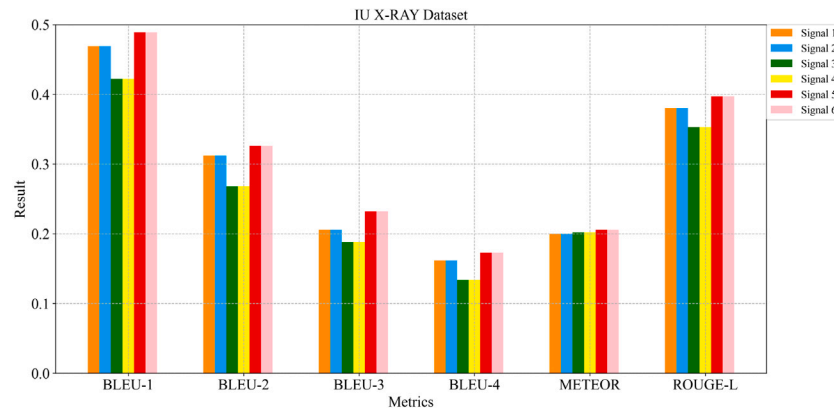


Fig. 7. Influence of different auxiliary signals on the experimental results.

Extensive experimental analysis shows that the method proposed in this paper outperforms other current network models. In the future, we will further reduce the complexity of the model and increase its inference speed while ensuring its effectiveness. Methods for selecting the auxiliary signal that is most suitable for the dataset will be another focus of our work.

#### CRediT authorship contribution statement

**Youyuan Xue:** Conceptualization, Methodology, Software, Writing – original draft. **Yun Tan:** Validation, Writing – review & editing, Project administration, Funding acquisition. **Ling Tan:** Data curation, Investigation. **Jiaohua Qin:** Supervision. **Xuyu Xiang:** Software, Visualization.

#### 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.

#### Data availability

The code has been released at <https://github.com/shangchengLu/ASGMN>.

#### Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62002392); in part by the Key Research and Development Plan of Hunan Province (No. 2019SK2022); in part by the Natural Science Foundation of Hunan Province, China (No. 2022JJ31019).

#### References

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE conference on computer vision and pattern recognition* (pp. 6077–6086). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2018.00636>, URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Anderson\\_Bottom-Up\\_and\\_Top-Down\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html).

- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. (pp. 4–5). <http://dx.doi.org/10.48550/arXiv.1607.06450>, ArXiv Preprint arXiv:1607.06450.
- Bae, J.-W., Lee, S.-H., Kim, W.-Y., Seong, J.-H., & Seo, D.-H. (2022). Image captioning model using part-of-speech guidance module for description with diverse vocabulary. *IEEE Access*, 10, 45219–45229.
- 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). Ann Arbor, Michigan: Association for Computational Linguistics, URL: <https://aclanthology.org/W05-0909>.
- 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). Online: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/2021.acl-long.459>, URL: <https://aclanthology.org/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). Online: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/2020.emnlp-main.112>, URL: <https://aclanthology.org/2020.emnlp-main.112>.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics, <http://dx.doi.org/10.3115/v1/D14-1179>, URL: <https://aclanthology.org/D14-1179>.
- Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304–310.
- Dognin, P. L., Melnyk, L., Mroueh, Y., Ross, J., & Sercu, T. (2019). Adversarial semantic alignment for improved image captions. In *IEEE conference on computer vision and pattern recognition* (pp. 10463–10471). Computer Vision Foundation / IEEE, <http://dx.doi.org/10.1109/CVPR.2019.01071>, URL: [http://openaccess.thecvf.com/content/CVPR\\_2019/html/Dognin\\_Adversarial\\_Semantic\\_Alignment\\_for\\_Improved\\_Image\\_Captions\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content/CVPR_2019/html/Dognin_Adversarial_Semantic_Alignment_for_Improved_Image_Captions_CVPR_2019_paper.html).
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15–29). Springer.
- Harzig, P., Chen, Y., Chen, F., & Lienhart, R. (2019). Addressing data bias problems for chest X-ray image report generation. In *30th British machine vision conference 2019* (p. 144). BMVA Press, URL: <https://bmvc2019.org/wp-content/uploads/papers/1007-paper.pdf>.
- 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). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2016.90>.
- Huang, L., Wang, W., Chen, J., & Wei, X. (2019). Attention on attention for image captioning. In *2019 IEEE/CVF international conference on computer vision* (pp. 4633–4642). IEEE, <http://dx.doi.org/10.1109/ICCV.2019.00473>.
- Jiang, W., Li, Q., Zhan, K., Fang, Y., & Shen, F. (2022). Hybrid attention network for image captioning. *Displays*, Article 102238.
- 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). Florence, Italy: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/P19-1657>, URL: <https://aclanthology.org/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 (volume 1: long papers)* (pp. 2577–2586). Melbourne, Australia: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/P18-1240>, URL: <https://aclanthology.org/P18-1240>.
- 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. (2019). MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). <http://dx.doi.org/10.13026/8360-t248>, PhysioNet.
- Kaur, N., & Mittal, A. (2022). CADxReport: Chest X-ray report generation using co-attention mechanism and reinforcement learning. *Computers in Biology and Medicine*, 145, Article 105498.
- Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 3337–3345). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2017.356>.
- Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *The 24th IEEE conference on computer vision and pattern recognition* (pp. 1601–1608). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2011.5995466>.
- Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In M. Balcan, & K. Q. Weinberger (Eds.), *JMLR workshop and conference proceedings: vol. 48, Proceedings of the 33rd international conference on machine learning* (pp. 1378–1387). JMLR.org, URL: <http://proceedings.mlr.press/v48/kumar16.html>.
- Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: annual conference on neural information processing systems 2018* (pp. 1537–1547). URL: <https://proceedings.neurips.cc/paper/2018/hash/e07413354875be01a996dc560274708e-Abstract.html>.
- Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence* (pp. 6666–6673). AAAI Press, <http://dx.doi.org/10.1609/aaai.v33i01.33016666>.
- Li, M., Liu, R., Wang, F., Chang, X., & Liang, X. (2022). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 1–18.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics, URL: <https://aclanthology.org/W04-1013>.
- 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).
- Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., & Jurafsky, D. (2021). Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 5288–5304). Online: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/2021.naacl-main.416>, URL: <https://aclanthology.org/2021.naacl-main.416>.
- 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). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, <http://dx.doi.org/10.3115/1073083.1073135>, URL: <https://aclanthology.org/P02-1040>.
- Qin, H., & Song, Y. (2022). Reinforced cross-modal alignment for radiology report generation. In *Findings of the association for computational linguistics* (pp. 448–458).
- Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28: annual conference on neural information processing systems 2015* (pp. 91–99). URL: <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>.
- 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. 1179–1195). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2017.131>.
- Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28: annual conference on neural information processing systems 2015* (pp. 2440–2448). URL: <https://proceedings.neurips.cc/paper/2015/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html>.
- Sun, C., Gan, C., & Nevatia, R. (2015). Automatic concept discovery from parallel text and visual corpora. In *2015 IEEE international conference on computer vision* (pp. 2596–2604). IEEE Computer Society, <http://dx.doi.org/10.1109/ICCV.2015.298>.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE conference on computer vision and pattern recognition* (pp. 3156–3164). IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2015.7298935>.
- Wang, C., & Gu, X. (2022). Image captioning with adaptive incremental global context attention. *Applied Intelligence*, 52(6), 6575–6597.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In F. R. Bach, & D. M. Blei (Eds.), *JMLR workshop and conference proceedings: vol. 37, Proceedings of the 32nd international conference on machine learning* (pp. 2048–2057). JMLR.org, URL: <http://proceedings.mlr.press/v37/xuc15.html>.
- Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In *International conference on medical image computing and computer-assisted intervention* (pp. 457–466). Springer.
- Yang, S., Wu, X., Ge, S., Zhou, S. K., & Xiao, L. (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, Article 102510.
- Yang, Y., Yu, J., Jiang, H., Han, W., Zhang, J., & Jiang, W. (2022). A contrastive triplet network for automatic chest X-ray reporting. *Neurocomputing*, 502, 71–83.
- 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, no. 07 (pp. 12910–12917).
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07 (pp. 13041–13049).
- Zhou, D., Yang, J., & Bao, R. (2022). Collaborative strategy network for spatial attention image captioning. *Applied Intelligence*, 52(8), 9017–9032.