Journal of the Franklin Institute 363 (2026) 108328

Contents lists available at ScienceDirect

Journal of the Franklin Institute

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

Bridging Visual Analysis and Text Generation: A Hierarchical Multi-scale Visual Feature Flow Model for Accessible Radiographic Report Automation

Hailong Zuo [a,1], Zhi Weng [a,1,*], Yunlu Duan [b], Zijing Lv [a], Feifan Bi [a], Zhiqiang Zheng [a]

[a] School of Electronic Information Engineering, Inner Mongolia University, Hohhot, 010021, China
[b] School of Microelectronics and Communication Engineering, Chongqing University, 401331, China

ARTICLE INFO

Keywords:
Cross-modal Alignment
Image-to-Text
Automatic Report Generation
Transformer
Natural Language Processing

ABSTRACT

Medical report generation is a currently popular research direction that combines image analysis and natural language processing technologies. It aims to relieve the pressure on doctors in writing reports and provide a diagnostic reference for them. Although current report generation technologies have made significant progress, there are still some problems. For example, the extraction of local lesion areas in medical images is insufficient, cross-modal alignment is difficult, and the extracted visual features are single-angled. Considering these issues, we propose the Hierarchical Multi-scale Visual Feature Flow (HMVF) model for medical report generation. Firstly, we introduce three feature extraction branches from different angles: disease label features, lung border features, and local lesion information features. This endows the model with the ability to focus on features from different perspectives of the image. To our knowledge, this is the first time that lung border features have been introduced into medical report generation for the MIMIC-CXR dataset. Then, the contents extracted from each branch are successively transmitted to the subsequent branches to achieve information interaction. At the same time, according to the characteristics of these visual features from different angles, we enhance the visual features of each branch respectively to prevent irrelevant information from diluting key information. Finally, we transmit the information at each level to different layers of the cross-modal information fusion module to generate the final text report. Extensive experiments on authoritative datasets demonstrate that our method outperforms others across multiple language evaluation metrics. On the IU-Xray dataset, our model scored 0.512, 0.365, 0.181, and 0.410 for BLEU-1/2/4, and ROUGE respectively. On the MIMIC-CXR dataset, scores were 0.391, 0.239, 0.161, and 0.115 for BLEU-1/2/3/4. These results confirm that the multi-branch feature extraction and hierarchical cross-modal fusion of HMVF effectively address the limitations of existing methods, providing a more robust solution for automated medical report generation. The model's design also offers a reference for multi-perspective feature utilization in other cross-modal generation tasks. Source code is available here.

* Corresponding author.
E-mail address: wzhi@imu.edu.cn (Z. Weng).
[1] Contribute equally to the article.

https://doi.org/10.1016/j.jfranklin.2025.108328
Received 8 August 2025; Received in revised form 14 November 2025; Accepted 13 December 2025
Available online 18 December 2025

1. Introduction

In recent years, the automatic generation technology of medical imaging reports has witnessed remarkable progress and has been integrated into the diagnostic platforms of certain hospitals, thereby assisting doctors in making diagnostic decisions [1,2]. Prior to the advent of this technology, doctors had to manually compose reports. This task not only required radiologists to possess extensive professional medical knowledge but also demanded that they accurately and concisely document the disease information contained in the images on the report forms. Especially in the current context where the number of patients in the radiology department is increasing steadily, radiologists are under immense pressure. They are burdened with the task of writing medical reports for extended periods at a high intensity [3–5]. This not only leads to fatigue and a decline in diagnostic efficiency but, more seriously, may even result in missed diagnoses and misdiagnoses. In such circumstances, the medical imaging report generation technology has promptly attracted the attention and participation of numerous artificial intelligence and medical researchers.

Currently, there exist a plethora of methods for generating medical imaging reports. These methods commonly adopt an end-to-end training paradigm, leveraging pre-trained CNN/VIT(Vision Transformer) networks as encoders to extract visual features [6,7]. Subsequently, the visual features and text embeddings are cross-modally aligned through a decoder. Therefore, the processes of extracting image features and cross-modal alignment are of paramount importance for the high-quality automatic generation of medical reports. Although these methods have achieved certain advancements, they still have several limitations. Firstly, when extracting visual features, there is insufficient attention paid to local regions of medical images, such as tissues, organs, or lesion areas. Secondly, the ability to perform cross-modal alignment with text embeddings is limited. Thirdly, the extracted visual features are single-angled, yet it is necessary to consider global and local aspects as well as the relevance to text embeddings simultaneously [8]. Evidently, these limitations increase the difficulty of achieving high-quality report generation.

As depicted in Figure 1, medical images contain a wealth of crucial descriptive information, including organ contours and symptom manifestations. If an anchor point in the image (which could be the neck, heart, lungs, or other locations) is identified, all other positions in the image can be recognized based on the orientation of this anchor point, enabling targeted descriptions. However, traditional automatic report generation methods overlook the integration of feature extraction from different angles with anchor points. This oversight leads to insufficient information extraction, such as inadequate extraction of global information or insufficient attention to local lesion areas, thus laying potential risks for subsequent cross-modal fusion.

In addition, in the scenario where an image is disease-free, traditional medical report generation models still necessitate the extraction of comprehensive image information. However, during the cross-modal alignment process, an excessive amount of information is not requisite. Despite this, the encoder component still has to convey a substantial volume of image information to the cross-modal fusion module. This significantly elevates the complexity of information processing for the encoder. (Specifically, the encoder is required to preserve the processing procedures for both diseased and healthy images. Given the limited parameters of the encoder, this undoubtedly poses a formidable challenge.) Consequently, a methodology that concurrently takes into account the extraction of disease categories, lung boundaries, and local lesion areas can effectively guide the model to acquire more diverse and rational visual features. Moreover, it supplies adequate and appropriate information for subsequent cross-modal fusion.

Taking the above-mentioned perspectives into consideration, we innovatively put forward a multi-level visual feature flow approach. This method integrates visual features from diverse angles to enhance balanced attention to all facets of medical images. Additionally, it conducts multi-angle visual-text alignment to boost the model's capacity for generating reports. For healthy images, a dedicated pathway is established to directly access the cross-modal information fusion module. This effectively reduces the complexity of information processing in other feature extraction phases. As a result, different branch components can concentrate more on their respective focused functions without being distracted. The principal contributions of this work can be encapsulated as follows:

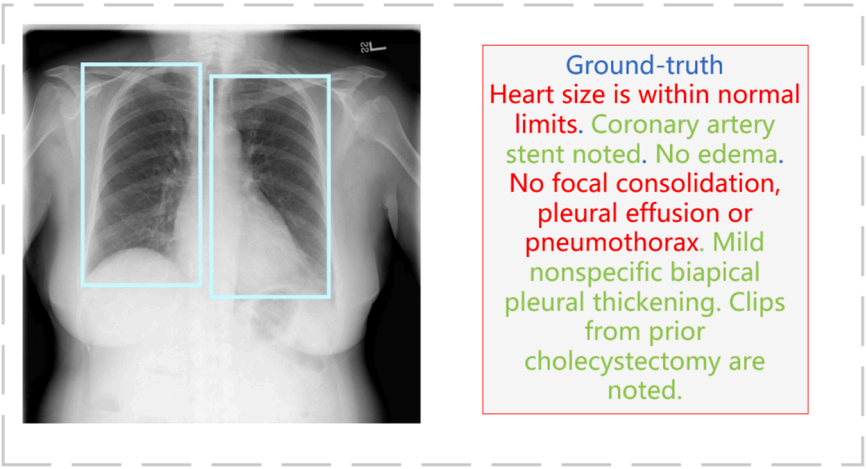


Fig. 1. Description of medical image annotations: The red color represents the descriptions of entities or organs, and the green color represents the descriptions of symptoms.

- For the first time, we incorporate the concept of lung anchors into medical reports within the MIMIC-CXR dataset, aiming to enhance the text report generation capabilities.
- We have designed a streamlined image category determination network. With fewer parameters, it can achieve performance comparable to that of the EfficientNet network in the same application scenario.
- By taking into account the distinct attributes of visual features, we align visual features from various angles with the text embeddings of medical reports. This ensures that information at different scales is accurately and comprehensively considered, thereby enhancing the reasoning ability of the network.
- We introduce a text report generator based on the Transformer-Decoder architecture and adopt a hierarchical cross-modal alignment strategy.

The subsequent structure of this article is as follows:

Section 2 delves into the related work and the challenges encountered. Section 3 offers a detailed account of the structure and formulas of HMVF. Section 4 presents the relevant theoretical analysis. Section 5 showcases the experimental results and their analysis. Section 6 provides discussions and limitations. Section 7 concludes the article with a summary and outlines future research directions.

2. Related Work

The task of generating medical imaging reports has its roots in image description. Consequently, this process primarily comprises two key aspects: the visual feature encoding process and the cross-modal alignment and decoding process. These two aspects are fundamental to the automatic generation of medical imaging reports, with the former focusing on extracting and representing visual information from medical images, and the latter aiming to align visual features with text embeddings and decode them into meaningful reports, thereby bridging the gap between the visual and textual modalities in the medical context.

2.1. Visual Feature Coder for Medical Images

The visual encoder for medical images represents the initial and crucial stage in the medical image report generation task. The outcome of visual feature extraction at this stage significantly influences the quality of the final medical report. In the past, the generation of medical image reports predominantly relied on integrated visual feature extraction. That is, a single extraction was performed on medical images, and the resultant features incorporated global information, local information, as well as descriptions of lesions or organs. TranSQ [9] proposed a Semantic Query learning paradigm. This paradigm aims to learn an intention embedding set and conduct a semantic query on visual features. To address the issues of missing internal edge features and difficult cross-modal data alignment [10], introduced a simple yet effective region-guided report generation model. In this model, the encoder detects anatomical regions and then describes a single prominent region to form the final report [11]. put forward a dual-modal visual feature flow (DMVF) for generating medical reports. They introduced region-level features based on grid-level features to enhance the ability of the encoder in this method to identify lesions and key regions [12]. proposed a knowledge-enhanced radiology report generation method. By introducing general knowledge and specific knowledge, they provided fine-grained knowledge for chest X-ray report generation [13]. incorporated contrastive learning as an auxiliary task for image feature learning and presented a multi-granularity report generation framework that combines sentence-level image-sentence contrastive learning. This framework can effectively learn knowledge from image-report pairs without the need for any additional labels [14,15]. The Token Mixer [16] framework contains an image encoder. It enhances cross-mode alignment by matching image-to-text generation with text-to-text generation that is less affected by exposure bias [17]. embedded the auxiliary image-text matching target into the encoder-decoder structure of the Transformer. This enables the learning of better-correlated image and text features, facilitating the report in differentiating similar images [18]. proposed an enhanced two-stage medical report generation model. This model combines medical images, clinical history, and writing style to jointly strengthen the encoder's capabilities [19]. introduced the Aggregate Discriminative Attention Map (ADM). It creates discriminative region maps using weak supervision, highlighting key regions to strengthen the ability to extract semantic information from medical images.

2.2. Semantic embedded decoder for medical image report

In the decoding stage of medical report generation [20], constructed a knowledge base capable of automatically extracting and restoring medical knowledge from text embeddings. During the decoder stage, the useful information in the knowledge base was integrated with the text information of the report [21]. proposed variational topic inference. By aligning visual and language patterns in the latent space, topics were inferred within a conditional variational inference framework. Each topic controlled the generation of a sentence in the report to improve the accuracy of the decoding stage. ATAG [22] introduced a novel knowledge graph structure composed of interconnected abnormal nodes and attribute nodes, which could better capture finer-grained abnormal details. This model also adopted an encoder-decoder architecture for report generation. To generate medical reports more effectively in the decoding stage [23], inserted a U-connection pattern between the encoder and the decoder to model the interactions between different modalities. Additionally, they developed symptom maps and knowledge distillation injectors to assist in report generation [24]. first extracted specific knowledge from retrieved reports and then added extra nodes or re-defined their relationships in a bottom-up manner, laying the foundation for subsequent text decoding processes. PPKED [25] explored prior knowledge from previous medical knowledge graphs (medical knowledge) and previous radiology reports (work experience) to mitigate text data bias in the decoding

stage.

Despite the significant progress made by the above-mentioned approaches, there are still limitations in methods related to feature extraction and cross-modal data alignment. Specifically, current methods mainly rely on the attention mechanism to learn visual features of medical images. This makes it challenging for the attention mechanism to capture visual information from different perspectives. The attention mechanism is unable to achieve balanced extraction when it comes to global information extraction, lesion boundary demarcation, and visual semantic extraction of lesion areas. Moreover, most existing methods currently extract different aspects of medical images as a whole. This makes it difficult to effectively align the positions of lesions or organs with text embeddings. In the decoding stage, only the overall visual features extracted are decoded in a one-time, non-hierarchical manner.

In view of the limitations of existing methods, we innovatively propose an automatic medical report generation framework named HMVF (Hierarchical Multi-scale Visual Feature Flow). The implementation details of this method are as follows: First, we utilize diverse backbone networks to separately extract the disease label features, regional boundary features, and lesion area features of medical images. These features respectively correspond to the global key information of the image, the lung bounding box, and more detailed disease information within the lesion area, jointly forming a multi-scale feature representation system. Subsequently, to further optimize these features, we meticulously designed two Feature Flow Adjustment Networks (FFAN). These networks can refine and enhance information based on the feature attributes extracted by their respective backbone networks, thus significantly improving the understanding ability of subsequent algorithms regarding the in-depth content of medical images. Finally, we constructed a Cross-Model Report Generation (CRG) module. This decoder can effectively integrate the adjusted visual feature streams from multiple branches mentioned above and precisely align them with the text embeddings extracted from medical reports. This process not only promotes the deep fusion of visual and text information but also enhances the expressiveness and accuracy of the algorithm in generating medical reports.

3. The proposed HMVF method

In this section, we will provide a detailed introduction to the HMVF method for medical report generation. The overall architecture of HMVF is shown in Figure 2 below.

3.1. Multi-level Visual Feature Flow Extractor

This method aims to extract visual information from medical images at multiple levels from three different perspectives. Next, we will introduce the visual information extraction processes at different levels in sequence.

Disease Label Extractor (DLE CLA): This extractor is designed to extract the disease labels contained in medical images, that is, to identify which diseases are present, without focusing on the location of the diseases or other contextual information. In previous studies, researchers often conducted this extraction process together with other extraction-perspective processes. However, this

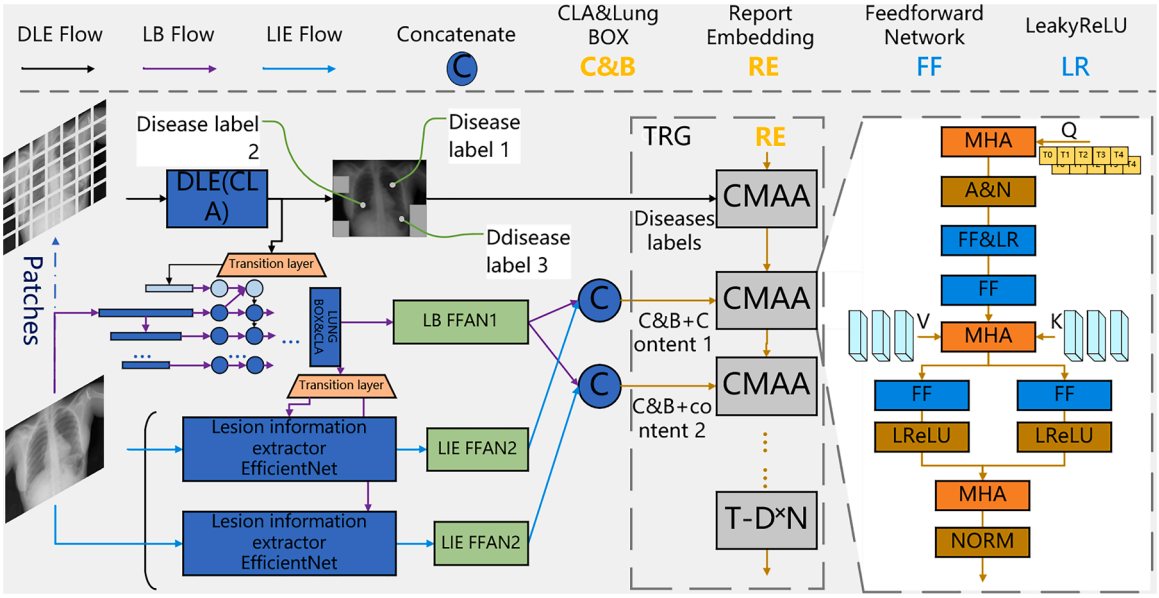


Fig. 2. Overview of the HMVF Method. The model is divided into three parts. The first part has a total of three feature streams, namely the Disease Label Extractor (DLE), Lung Box Extractor (LBE), and Lesion Information Extractor. The Lesion Information Extractor contains two identical branches. The second part is the Feature Flow Adjustment Network (FFAN) for each feature stream branch. The last part is the Text Report Generator (TRG).

approach increases the uncertainty of extraction and may lead to overlooking certain diseases. Additionally, Vision-Transformers were typically used to extract features from patched images. To proactively enhance the model's attention to disease information and considering the network structures of the boundary locator and the lesion information extractor, we independently designed a disease label extractor with MBConv (Mobile inverted convolution). This ensures that the model will definitely take disease category information into account and significantly reduces the model parameters compared to the complete efficientnet network.

For a given medical image ᑭ, ᑭ is first divided into a specified number of fixed-sized patches (10 × 10). Convolution operations are then applied to these patches respectively, followed by pooling operations on the results of the convolution. These pooled results are used in the Elimination-Multi-Head Attention layer (E-MHA) to calculate scores for each patch within the range of (0,1). Finally, the scores are multiplied by the corresponding patch regions, and then MBConv convolutions are continuously applied until the labels of each disease are finally output. This process can be regarded as a form of patch-information enhancement, where relatively unimportant patches are weakened while important ones are strengthened. The final generated result can also be considered as a form of contextual cueing information, indicating which diseases the medical image has. The process is expressed by the following formulas:

$$I_i^M = E - MHA(P(Conv(I_i))) \tag{1}$$

E-MHA represents the Elimination-Multi-Head Attention layer, P represents the adaptive average pooling operation, Conv represents convolution, and I_i^M represents the i-th patch in image ᑭ after passing through the E-MHA layer. Subsequently, multiple MBConv operations are performed on I_i^M, and finally, the labels of each disease are output through an n + m classifier, where n is the number of disease labels and m is the label redundancy. The formula is expressed as follows:

$$cla\{cla_1, cla_2, \ldots, cla_n\} = Cla\left(MBConvn * N\left(I_i^M\right)\right) \tag{2}$$

Here, cla represents a classifier with two layers. $D^{1\times(n+m)}$ is the dimension of cla. In this context, n represents the ratio of the number of output channels to that of input channels, N represents the number of MBConv in each layer, and cla_n represents the obtained disease labels. Subsequently, cla_n is respectively transmitted to the first layer of the report generator and the boundary locator layer to provide cues about the potential disease-related contextual information contained in the medical image.

For further information, please refer to Figure 3 below.

Lung Border Extractor (LBE): Considering the close relationship between boundary determination and content analysis, we employed the EfficientNet network for lesion information extraction. For the determination of lung borders, we utilized the EfficientDet network, which is developed based on the former. We hypothesized that visual features extracted by similar networks are more likely to be utilized by each other.

The EfficientDet receives inputs from two different sources to detect the lung region and then transmits the output to the subsequent lesion information extractor and the cross-modal report generator. The formula is described as follows:

$$I^b = Effd(I, tran(cla)) \tag{3}$$

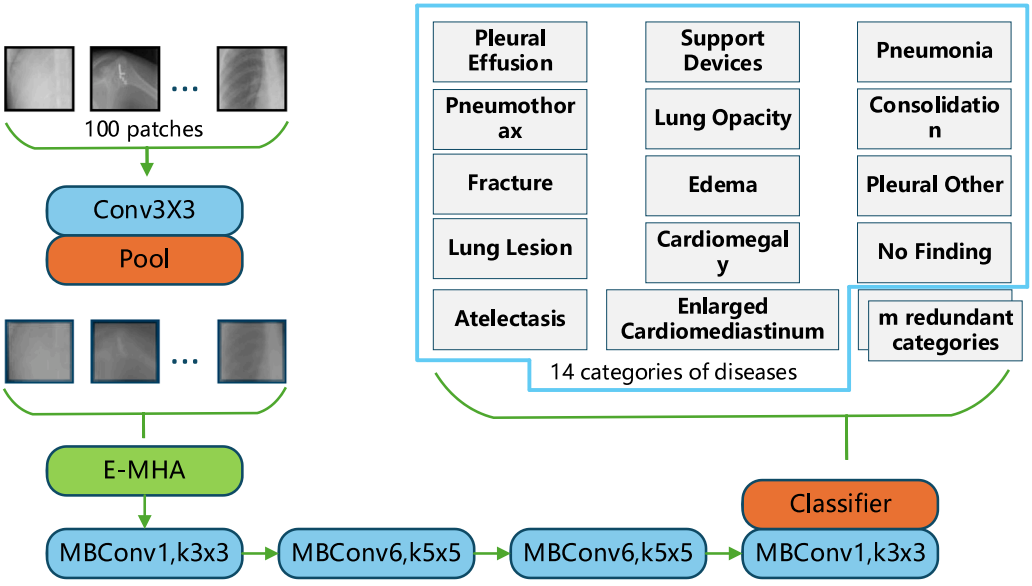


Fig. 3. Based on the improvement of the EfficientNet network, two of the disease labels are relatively special, namely "Support Devices" and "No Finding". Since these two labels are classified as disease categories in most medical image report generation technologies, we also classify them into disease categories here for the sake of consistency and to facilitate reading for subsequent researchers. Therefore, when these two special labels are included, the total number of disease labels amounts to 14.

Here, *Effd* represents the modified EfficientDet network. Specifically, a layer for receiving disease labels is added, and the last layer for class prediction is removed, as shown in Figure 2. $\text{ᑭ}^{b}=\{\text{ᑭ}_1^{b}, \text{ᑭ}_2^{b}, ..., \text{ᑭ}_n^{b}\}\epsilon R^{D}$, representing the results of lesion boundary extraction, where D is the dimension of ᑭ_i^{b}. ᑭ represents the original image without patch processing, and tran represents a transition layer composed of an MHA (Multi-Head Attention) and two linear layers. The formula is described as follows:

$$tran = L_2(A_{swish}(L_1(MHA(in)))) \tag{4}$$

L represents the linear layer, MHA represents the multi-head attention mechanism, and there is a swish activation function between the two linear layers. This boundary locator branch only extracts information such as the lung region or position, as shown in Figure 4. Then, I^{b} is used in LBE FFAN1 and is respectively combined with the outputs of the two lesion information extractors and fed into different layers of the text generator to further enhance the ability of the text generator.

Lesion Information Extractor (LIE): The lesion information extractor conducts in-depth feature extraction of the lesion area based on disease label and lung position information, which can effectively focus on key lesions. Here, we deploy two EfficientNet network extractors with the same structure to extract lesion information at different depths or dimensions. These two extractors also receive inputs from two different sources: the medical image and the output of the lung border extractor from the previous layer.

For a medical image ᑭ, it will be appended with information from the boundary locator that contains disease label information, and then transmitted to the EfficientNet network for targeted visual feature analysis. The formula is described as follows:

$$I = I \oplus Tran(I^{b}) \tag{5}$$

Here, $\oplus$ represents an operation performed on ᑭ, based on the results provided by I^{b}, which enhances the pixel values therein. The formula is described as follows:

$$I_i^{c} = Effn(I)i = 1, 2 \tag{6}$$

Among them, Effn represents the EfficientNet network, and $I_i^{c}=\left\{I_{i_1}^{c}, I_{i_2}^{c}, \ldots, I_{i_n}^{c}\right\} \in R^{D}$ represents the output result of the i-th lesion area information extractor, where D is the dimension of $I_{i_n}^{c}$.

cla, I_b, I_i^{c} can be regarded as feature extractions from three different perspectives of the image. Each of the extracted features not only exerts an impact on other branches but also has the opportunity to transmit information backward to the report generator. This significantly enhances the influence of various information dimensions at different layers of the report generator.

3.2. Feature Flow Adjustment Network

The extracted visual features, except for those from the disease label branch, will be adjusted to varying degrees to better meet the requirements of the report generator. The FFAN network, as shown in Figure 2, consists of two adjustment modules with different structures, which are respectively used for the lung border extractor and the lesion information extractor.

LBE FFAN1: The boundary information enhancer is based on MHA. However, different from MHA, it introduces a pooling operation and a linear layer after Softmax. Moreover, to mitigate the insufficient information extraction by the attention mechanism, a linear-layer branch with an activation function is separately introduced to enhance the information correlation within the same lesion area, as shown in Figure 2.

The boundary information enhancement layer will apply linear-layer transformation to each feature separately to generate the V, K, and Q feature maps. Then, the V and K feature maps undergo a Hadamard product operation, followed by global pooling. The result is multiplied by the globally-pooled K, and then the weights are calculated through Softmax. The result is added to the V feature map, and finally, calculations are carried out through a linear layer. The formula explanation is as follows:

$$Q = L_q(I^{b}) \tag{7}$$

$$K = L_k(I^{b}) \tag{8}$$

$$V = L_v(I^{b}) \tag{9}$$

Q, K, and V represent the query, key, and value features of the image respectively, and L represents the corresponding linear transformation, as explained in Eq. (3).

$$w_1 = soft(P(Q_1 \cdot K_1), P(Q_2 \cdot K_2), \ldots, P(Q_n \cdot K_n)) \tag{10}$$

Among them, w_1 is the weight value of each piece of information, and Q_n, K_n are different query and key matrices of the same image.

$$V_n = V_n * W_n \tag{11}$$

Here, $V_n \in R^{D} = I^{b}$ is the feature data to be fed into the subsequent linear layer. Finally, this branch can be completed by passing through a linear layer with a LeakyReLU activation function, as shown in the following formula.

$$V_1 = LR(L(V_n)) \tag{12}$$

LR represents the LeakyReLU activation function. After completing the first branch, a concat operation needs to be performed

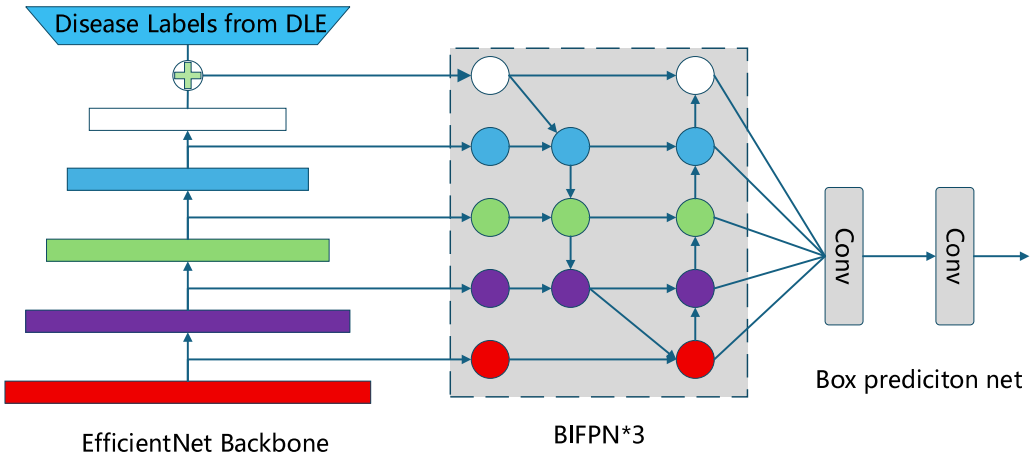


Fig. 4. We have made some minor modifications to the EfficientDet network. First, we added a disease label input, as shown in the "labels from DLE" section in the figure. Here, we adopted a simple concatenation method to fuse the bounding box information with the disease label information, and thus the dimension of the subsequent fusion vector increases accordingly. Second, the final category information extraction component has been removed and replaced with a separate bounding box prediction network.

between V_1 and V_2 from the second branch. The second branch is mainly designed to prevent insufficient information transmission caused by the attention mechanism paying too little attention to certain features. The relevant formulas are as follows:

$$V_{2_1} = LR\left(L\left(I^b\right)\right) \tag{13}$$

$$V_2 = LR(L(V_{2_1})) \tag{14}$$

$$I^B = V_1 + V_2 \tag{15}$$

Here, I^B is the output of the boundary information enhancement layer. Subsequently, it will be combined with two lesion information extractors of the same structure to jointly provide visual feature information for the second layer of the report generator.

LIE FFAN2: Compared with the boundary information enhancement layer, the structure of the lesion area information enhancement layer is relatively simpler, which merely performing the MHA operation on their respective dimensions.

Considering the powerful transformation ability of the fully-connected layer, the output of the EfficientNet network undergoes two linear transformations (the first layer uses the swish activation function, and the second layer uses the sigmoid activation function), and then by applying the MHA of the Transformer, the respective weights can be obtained. The formula is described as follows:

$$S_i = MHA\left(L_1\left(I_{i_1}^C\right), L_2\left(I_{i_1}^C\right), \ldots\right) \tag{16}$$

In the formula, S_i represents the score of the i-th lesion information extractor, and $I_{i_n}^C$ represents the n-th dimensional feature of the i-th lesion generator.

Then, multiply S_{i_n} by the respective $I_{i_n}^C$, and thus the weighted lesion area information is obtained. The formula is described as follows:

$$I_i^C = I_{i_n}^c \cdot S_{i_n} \tag{17}$$

I_i^C is the output of the information enhancement layer on the branch of the i-th lesion information extractor. In our network, there are two lesion information extractors with the same structure responsible for extracting information from different perspectives.

3.3. Cross-Modal Report Generator

We have extracted visual features from medical images from three different perspectives: disease labels, lung borders, and lesion information in medical images. To effectively utilize these features, we have designed a multi-level cross-modal report generator. The first three layers are respectively used for attention alignment with visual features at different scales, and the subsequent layers adopt the decoder structure in the standard Transformer. This generator can effectively align visual features from different angles with text features in a hierarchical manner, thereby generating high-quality medical imaging reports.

Firstly, we conduct Q, K, V calculations on the information from disease labels and text information. In order to align the data between different modalities, we exchange the q of the disease labels and the q of the text information, and then calculate the Multi-Head Attention (MHA). The formula is expressed as follows:

$$X_1^t = MHA_1(Q_l, K_t, V_t) \tag{18}$$

$$X_1^l = MHA_1(Q_t, K_l, V_l) \tag{19}$$

The superscript in X_n^m represents the text or label, and the subscript represents the n-th step of calculation for cross-modal attention alignment. After obtaining X_n^m, a residual connection and a normalization operation are then carried out with the original cross-modal attention alignment. Considering the fully-connected expressive ability of the linear layer, we then conduct two feed-forward neural network calculation steps. The relevant formulas are shown as follows:

$$X_2^t = X_1^t + W_e \tag{20}$$

$$X_2^l = X_1^l + D_l \tag{21}$$

$$X_4^t = L(L(X_2^t)) \tag{22}$$

$$X_4^l = L(L(X_2^l)) \tag{23}$$

Among them, W_e and D_l represent the word embedding and the disease label information flow respectively. So far, we have completed the first MHA operation among the three attentions in cross-modal attention alignment. Then, it is followed by two linear layers with LR activation function branches, which are used to enhance its nonlinear ability. After that, another MHA is performed and then a normalization operation is carried out, thus completing the calculation of a complete cross-modal attention alignment layer. The formula is described as follows:

$$X_8 = Norm(MHA_3(L_1(MHA_2(X_4^t, X_4^l)), L_2(MHA_2(X_4^t, X_4^l)))) \tag{24}$$

X_8 is the output of a layer in cross-modal attention alignment. Each of the first three layers absorbs one of the visual features refined from the medical report. Note that the CMAA modules of the second and third layers absorb visual features at the second MHA, which is different from the first layer. This process can be observed in Figure 2.

Furthermore, for the output X_8 of the third layer, it will be passed to several Transformer decoders again. Here we set three. After the last Transformer decoder, a linear classifier is connected to generate the final medical imaging report text (Algorithm 1).

4. Optimization Analysis of the Loss Function

In order to enhance the quality of medical report generation, we optimize the algorithmic inference by means of multiple alignment losses to facilitate the alignment among data from different modalities. These alignment losses include disease category label alignment, lung border feature alignment, and embedded text alignment. These three types of losses respectively correspond to the disease label extractor, the lung border extractor of HMVF, and the finally generated text report component.

Disease Label Feature Loss: To boost the ability of extracting disease label features, we take inspiration from word2vec and employ the visual-text triplet loss. This loss is applicable to medical report generation tasks that involve a large quantity of image-text paired data. The disease labels are common diseases that we extract from medical reports using CheXpert[2] Specifically, we calculate the differences between the images and the paired as well as unpaired labels, so as to guide the alignment of visual features with the correct disease labels and keep a distance from the incorrect labels. Thus, the extraction of grid visual features can be optimized. The formula is expressed as:

$$L_{cla} = \max\left(dis(cla, \overline{cla}) + \frac{\alpha}{dis(cla, \overline{\overline{cla}})}, 0\right) \tag{25}$$

$$dis(cla, \overline{cla}) = 1 - \cos(cla, \overline{cla}) = 1 - \frac{cla \cdot \overline{cla}}{|cla| \cdot |\overline{cla}|} \tag{26}$$

$\overline{cla}$ and $\overline{\overline{cla}}$ respectively represent the positive and negative labels. The negative labels are randomly sampled from outside the positive ones. "dis" represents the calculation of cosine similarity, and "α" represents the scaling hyperparameter, which is used to control the proportion of negative samples in the L_{cla} loss.

BLE Information Feature Loss: To enhance the ability of aligning the text report with the information of the lesion area, we introduce the BLE information feature loss. This loss is used to adjust the network's regional prediction ability. This loss function is primarily intended to empower the BLE branch to furnish the entire network with lung bounding box information, while propagating its own information in conjunction with that derived from the CLA branch to subsequent network segments. Since the GIoU loss function is particularly suitable for object detection tasks, we no longer use a loss function similar to that in the EfficientDet network. The relevant formula is as follows:

$$IoU = \frac{A \cap B}{A \cup B} \tag{27}$$

[2] https://github.com/MIT-LCP/MIMIC-CXR/tree/master/txt/CheXpert

Algorithm 1
Learning procedure of HMVF.

Input: medical images $I = \{I_1, I_2, \ldots, I_n\}$; Corresponding Medical Image Report $Y = \{Y_1, Y_2, \ldots, Y_n\}$
Output: generated Medical diagnosis report Y'
1:**for epoch in epochs do**
2: Sample batch medical images and their corresponding text reports.
3: **for** $image_i$ **in** I**:**
4: $clas_i$=**Disease Label Extractor** (image) {Eq. (2)}
5: I^b=information enhancement 1 (**LBE FFAN1** (image+tran ($clas_i$))) {Eq. (3)}
6: I_{c_i}=information enhancement 2 (**LIE FFAN2** (image+tran (I^b)) {Eq. (6)}
7: **for** L **in {the number of cross-modality attention layer}:**
8: $x = \mathrm{CAAL1}(clas_i, TE)$ {x represents the output of the first three layers of the report generator.}
9: $x = \mathrm{CAAL2}\left(x, \mathrm{concat}\left(I_{c_2}, I^b\right)\right)$
10: $x = \mathrm{CAAL3}\left(x, \mathrm{concat}\left(I_{c_1}, I^b\right)\right)$
11: transformer decoder*N (x)
12: $p\left(y_t^* | y_1 : y_n\right)$ {Generate medical reports word by word}
13: $L_{\mathrm{all}} \leftarrow \left(L_{\mathrm{CLA}}, L_B, p\left(y_t^* | y_1 : y_n\right)\right)$ {Three considerations of different scales for calculating the loss function}
14:**end for**

$$GIoU = IoU - \frac{C - (A \cup B)}{C} \tag{28}$$

$$L_B(\overline{x}, \overline{y}) = 1 - GIoU \tag{29}$$

A and B represent the ground-truth bounding box and the predicted bounding box respectively. C represents the area of the minimum enclosing rectangle that contains the two bounding boxes. A∪B represents the area of the union of the two bounding boxes. C−(A∪B) represents the area of the minimum enclosing rectangle that does not belong to the two bounding boxes. The value range of GIoU is [-1, 1]. When the two bounding boxes completely overlap, GIoU = IoU = 1; when the two bounding boxes are completely non-overlapping and far apart, GIoU tends to -1. $\overline{x}$ is the predicted lung bounding box information, and $\overline{y}$ is the real ground truth of the lung.

Text Report Generation Loss: In order to optimize the report generation process, we calculate the character-level cross-entropy loss between the generated text report and the real report. The formula is expressed as follows:

$$L_t = -\sum_{i=1}^{N} p_i \log(\widehat{p}_i(p_1, p_2, \ldots, pi-1)) \tag{30}$$

$\widehat{p}$ is the predicted character probability, and p_i is the text before time step i, which is used to predict the character-level probability of the character at time step i.

In order to comprehensively improve the performance of the algorithm, we integrate these three loss functions designed from different perspectives into a unified optimization objective. The specific form is shown in the following formula, so as to achieve the joint optimization of the overall algorithm.

$$L_{all} = \lambda_1 L_{cla} + \lambda_2 L_B + \lambda_3 L_t, \lambda_1 + \lambda_2 + \lambda_3 = 1, (\lambda_1 > 0, \lambda_2 > 0, \lambda_3 > 0) \tag{31}$$

Among them, λ is the contribution coefficient of each term in the formula to the overall L_{all}. The constraint that the sum of the coefficients is 1 ensures that these weights maintain a balance of "relative importance" during the allocation, preventing a certain loss term from dominating the training process due to an excessively large weight, and also avoiding being ignored due to an excessively small weight. In our training, the values of the above coefficients are 0.3, 0.3, and 0.4 respectively.

5. Experiments

To evaluate the effectiveness of the HMVF network, we conducted a large number of experiments on two widely used authoritative datasets. First, we provided a detailed introduction to the two datasets used in the experiments. Then, we described the experimental implementation details and the evaluation criteria for the generated reports. Furthermore, we compared HMVF with the state-of-the-art medical report generation models and presented both objective and subjective results. Finally, we analyzed the impact of visual features from different branches on medical report generation to verify the effectiveness of the proposed method.

5.1. Datasets

IU-Xray. The IU-Xray [26] Indiana University Chest X-Rays dataset is a valuable resource in the field of medical images and is used for research on the automatic generation of medical reports. It was collected by Indiana University and contains 7,470 preprocessed chest X-ray images (in PNG format) and 3,955 corresponding diagnostic reports. The reports cover sections such as comparison, indication, findings, and impressions. Among them, the findings and impressions are particularly crucial for report generation.

MIMIC-CXR. The MIMIC-CXR v2.0.0 dataset [27] contains 377,110 chest X-ray images in DICOM format from the Beth Israel Deaconess Medical Center in Boston and 227,835 corresponding radiology reports. The dataset has been de-identified, complies with

HIPAA requirements, and has removed PHI. It aims to support research in medical image understanding, natural language processing, and decision support. The segmented images in this dataset are sourced from the Chest X-ray Dataset with Lung Segmentation dataset [28].

To ensure the objective fairness of the performance comparison, we used the official dataset division standard for the MIMIC-CXR dataset. For the IU-Xray dataset, we adopted the division standard proposed by [29], which is train: 7, evaluation: 1, and test: 2.

5.2. Implementation Details

To ensure the normal progress of the training, we set the epoch and batch size to 35 and 32 respectively for the MIMIC-CXR dataset, and for other datasets, they are set to 60 and 8. The feature dimensions of the label extractor, the bounding box locator, and the lesion information extractor are 512, 1024, and 2048 respectively. The initial image resolution is 512×512. Since the bounding box information extractor and the disease label extractor require pre-training, and the MHVF is jointly trained after the pre-training is completed, the learning rates of these two branches are set to 1e-6. The learning rates of the two branches of the remaining lesion information extractor are set to 1e-5. We adopt a beam search strategy with a size of 3. In terms of data processing, we replaced all punctuation marks in the reports with commas. Only commas are generated during the generation stage. We also added <BOS> and <EOS> to indicate the start and end of each report, and set the words that appear less than 5 times as <UNK>.

5.3. Evaluation Methods

To verify the language fluency and clinical effectiveness of the proposed HMVF framework, we adopted several representative evaluation metrics, such as the BLEU [30], METEOR [31], and ROUGE [32] methods. For BLEU and METEOR, we used the open-source nltk library for implementation. To ensure a fair comparison, we pre-aligned the scores provided by each model to ensure the consistency of the evaluation results. For ROUGE, we adopted the pltrdy/ruge[3] codebase for implementation and carried out a similar comparison between its calculation results and the corresponding model scoring results to ensure accuracy.

BLEU is a machine translation evaluation metric based on n-gram matching. It is fast to calculate and has a low cost, but it ignores synonyms and sentence structures. METEOR is more comprehensive, taking into account precision and recall, as well as word order penalties, but it relies on external resources. ROUGE is used to evaluate summaries and translations, intuitively reflecting the degree of overlap, but it has a low discrimination degree and is not sensitive to lexical changes. Therefore, only when each metric can achieve a high score as much as possible can the HMVF method be considered effective.

5.4. Quantitative Analysis

Table 1 shows the comparison results between HMVF and other SOTA models on the IU-Xray dataset, and Table 2 presents the comparison results between HMVF and other models on the MIMIC-CXR dataset. These two comparison tables are used to demonstrate the advantages and effectiveness of the proposed HMVF method.

Previous models include classic image captioning models such as COATT [39], HRGR [44], ADAATT [34], ATT2IN [33], ST [37], and more recent medical report generation models such as SSVE [42], CAMA [41], CGA [40], PPKED [45], CMN [29], R2Gen [36], CMAS-RL [38]. These models involve technologies such as Transformers, knowledge bases, and CNNs, etc., and they can provide more comprehensive comparison results.

As shown in Table 1, compared with previous methods, our method has achieved significant improvements in the results and obtained the best performance in most of the evaluation metrics. Compared with the second-best model, HMVF has improved by 0.8%, 0.2%, 0.1%, and 0.2% in bleu1, bleu2, bleu4, and rouge respectively. However, the score of HMVF on meteor has not reached the level of the score of the CGA model. This situation may be attributed to the optimized reinforcement learning strategy adopted by the CGA model, which has shown strong effectiveness in cross-modal alignment. In Table 2, we have achieved the highest scores of 0.391, 0.239, 0.161, and 0.115 in bleu1, bleu2, and bleu4. Although the scores on meteor and rouge have not reached the highest, the scores obtained are still highly comparable to those of the CGA model with the highest score. Although CGA uses the strategy of explicitly quantifying the visual uncertainty and text uncertainty in radiology report generation, making its score performance more outstanding, our HMVF method outperforms the CGA method in more evaluation metrics (4 out of 6 evaluation metrics are better than the CGA results). Table 2 also shows a similar trend. It is worth noting that the NLG score in Table 2 is slightly lower than that in Table 1, which is also consistent with the situation that the MIMIC-CXR dataset has a higher image resolution and the medical report text is longer and more complex.

Overall, such excellent results have been achieved mainly because our model adopts a hierarchical extraction strategy in medical image feature extraction, and during the decoding stage, visual information at different scales is fed into different layers of the report generator. This enables different layers of the report generator to perform cross-modal alignment in a targeted manner and pass their results to the next layer for further use or as information clues.

[3] https://github.com/pltrdy/rouge

Table 1
Comparisons of the HMVF model with previous studies on IU X-RAY with NLG metrics, "/" represents no data.

IU-XRAY						
MODEL OR METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
ATT2IN [33]	0.224	0.129	0.089	0.068	/	0.308
ADAATT [34]	0.220	0.127	0.089	0.068	/	0.308
HRGR [35]	0.438	0.298	0.208	0.151	/	0.322
PPKED [25]	0.483	0.315	0.224	0.168	/	0.376
R2Gen [36]	0.470	0.304	0.219	0.165	0.187	0.371
ST [37]	0.216	0.124	0.087	0.066	/	0.306
CMN [29]	0.475	0.309	0.222	0.170	/	0.375
CMAS-RL [38]	0.464	0.301	0.210	0.154	/	0.362
COATT [39]	0.455	0.288	0.205	0.154	/	0.369
CGA [40]	0.497	0.357	0.279	0.225	0.217	0.408
CAMA [41]	0.504	0.363	0.279	0.218	0.203	0.404
SSVE [42]	0.492	0.321	0.233	0.180	0.203	0.379
Ours	**0.512**	**0.365**	0.231	**0.181**	0.212	**0.410**

Table 2
Comparisons of the HMVF model with previous studies on MIMIC-CXR with NLG metrics.

MIMIC-CXR						
MODEL OR METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
ATT2IN [33]	0.325	0.203	0.136	0.096	0.134	0.276
ADAATT [43]	0.299	0.185	0.124	0.088	0.118	0.226
PPKED [25]	0.360	0.224	0.149	0.106	0.149	0.284
ST [37]	0.299	0.184	0.121	0.084	0.124	0.263
R2Gen [36]	0.353	0.218	0.145	0.103	0.142	0.277
CMN [29]	0.353	0.218	0.148	0.106	0.142	0.278
CAMA [41]	0.374	0.230	0.155	0.112	0.145	0.279
SSVE [42]	0.373	0.230	0.161	0.112	0.162	0.297
Ours	**0.391**	**0.239**	**0.161**	**0.115**	0.160	0.293

5.5. *Clinical efficacy*

To more comprehensively evaluate the effectiveness of the model, we introduced the clinical effectiveness evaluation. Moreover, we specifically collected models with available source codes for comparison. As shown in Table 3, due to the adoption of the hierarchical multi-branch strategy for extracting visual features, our model has scores in precision, recall, and F1-score on the IU-Xray dataset that are much higher than those of other benchmark models. In the previous NLG (Natural Language Generation) evaluation, although the CMN model outperformed our HMVF model in some individual scores, in the clinical effectiveness evaluation, we achieved the best results in all evaluation metrics compared with the CMN model. Notably, our model achieves the same precision (0.431) as M2KT on MIMIC-CXR, but outperforms it in F1-score (0.412 vs. 0.406), which benefits from our hierarchical cross-modal alignment (Section 3.3). The score improvement on the IU-Xray dataset is higher than that on the MIMIC-CXR dataset, which is related to the fact that the IU-Xray dataset is smaller and easier to fit. The clinical effectiveness evaluation and the NLG evaluation jointly verify the superiority of the proposed HMVF method.

Table 3
HMVF Clinical Efficacy metrics on IU X-Ray and MIMIC-CXR datasets with Precision, Recall, F1-score.

Datasets	Model	Precision	Recall	F1-score
IU-Xray	MRAM [46]	0.341	0.251	0.250
	R2Gen [47]	0.353	0.278	0.281
	CMN [48]	0.356	0.282	0.288
	R2RL [49]	0.361	0.298	0.301
	M2KT [50]	0.316	0.308	0.292
	Ours	**0.388**	**0.367**	**0.361**
MIMIC-CXR	MRAM	0.332	0.238	0.243
	R2Gen	0.349	0.267	0.266
	CMN	0.351	0.273	0.274
	R2RL	0.359	0.294	0.297
	M2KT	0.431	0.342	0.406
	Ours	**0.431**	**0.371**	**0.412**

5.6. Qualitative Analysis

In addition to the objective comparison results, we also provide as many subjective comparison results as possible to measure the effectiveness of the proposed method, as shown in Figure 5 and Figure 6.

Since we use the lung segmentation information provided by the MIMIC-CXR dataset, the magnitude of weight updates for the bounding box extractor during the training of IU-Xray was very small. However, the final text report generation achieved results comparable to the ground truth.

The content of the report in Figure 5 is basically consistent with the ground truth. It accurately predicted the size of the heart, the location of the nodules, and the lesion conditions such as lung expansion, and there were no sentence-level errors. Such errors are common in many medical report generations, but we did not observe this situation in our report results, nor did we observe any missed diagnosis. This may be due to the better ability of our report generator to align visual features. In addition, the situation where the model pays more attention to positive samples and ignores negative samples due to the bias between positive and negative samples in the dataset was not significantly observed in our report generation. However, for many positive sample diagnosis situations, fixed expressions still appeared. This may be because there are many similar sentences in the healthy descriptions of positive samples. The generation results of both positive and negative samples in our report have reached a level comparable to that of the real reports. In the second image, there is an obvious error where "normal" was wrongly described as "abnormal", but the subsequent "normal" was not wrong. Therefore, we speculate that the visual features extracted the relevant information, but there was a problem during the cross-modal attention alignment. It is worth noting that there are many occurrences of the word "the" in our report. This is because the report generator has learned the usage methods and characteristics of "the" in language logic, which is a phenomenon not found in other medical report generation models.

Figure 6 shows the report generation results on the MIMIC-CXR dataset, and some trends similar to those in the IU-Xray dataset are observed. Even though the report texts in the MIMIC-CXR dataset are significantly more complex, our text generator can still generate the report content quite well. In the first image, CTA is omitted, which may be due to the fact that CTA appears too infrequently. In the third image, there is a misrepresentation where "periodic deterioration" is wrongly interpreted as "gradual deterioration". For the rest, they are all close to the content covered by the ground truth.

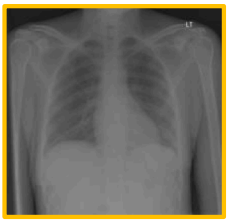
First, CTA is omitted primarily due to data scarcity: as highlighted in Section 6.2 (Discussion), medical report datasets suffer from

Ground truth

Hyperinflated lungs with no focal alveolar consolidation. No definite pleural effusion seen. Heart size near top normal limits, aortic calcifications and mild ectasia/tortuosity. No typical findings of pulmonary edema. Contour irregularity of the left clavicle appears chronic and suggests old injury.

Ours

the lungs are hyperinflated with no focal alveolar consolidation observed, there is no definitive evidence of pleural effusion, the heart size is near the upper limit of normal, accompanied by aortic calcifications and mild dilation/tortuosity, typical findings of pulmonary edema are absent, the contour irregularity of the left clavicle suggests a chronic condition, possibly an old injury,

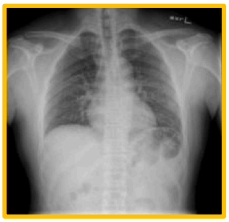


Ground truth

The cardiac silhouette is normal in size and configuration. The mediastinum and perihilar structures appear unremarkable. The lungs appear clear. The osseous structures are within normal limits.

Ours

the cardiac silhouette exhibits abnormal size and configuration, the mediastinum and perihilar structures appear normal without any remarkable findings, the lungs are clear with no abnormalities detected, the osseous structures are within normal limits,



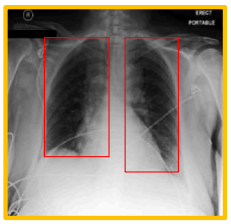
Ground truth

Heart size within normal limits. Small, nodular opacity in the right upper lobe. This does not look like an acute infiltrate, and more XXXX represents a granuloma. No pneumothorax or effusions,

Ours

the heart size is within the normal range, a small, nodular opacity is noted in the right upper lobe, which does not resemble an acute infiltrate and is more indicative of a granuloma, no signs of pneumothorax or pleural effusion are present,

Fig. 5. A comparison between the generated results on the IU-Xray dataset and the ground truth. Light green represents diagnostic results with the same meaning, red represents abnormal report results.



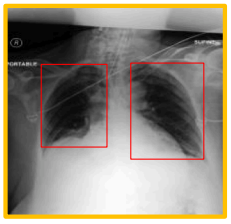
Atelectasis

Ground truth

There has been interval extubation and improved lung volumes compared to the recent radiograph. Bibasilar atelectasis has nearly resolved with residual patchy atelectasis remaining in the right lower lobe and only minimal residual linear atelectasis in the left lower lobe. Apparent rightward deviation of the trachea is likely due to mild patient rotation and curvature of the spine, as there is no evidence of a discrete paratracheal mass on recent neck CTA of ___. Cardiac silhouette is stable in size. No pleural effusion or pneumothorax.

Ours

recent extubation has led to improved lung volumes compared to previous radiographs, bibasilar atelectasis has almost resolved, with only residual patchy atelectasis in the right lower lobe and minimal linear atelectasis in the left lower lobe, the apparent rightward deviation of the trachea is likely due to mild patient rotation and spinal curvature, as no discrete paratracheal mass is evident on recent neck ~~CTA~~, the cardiac silhouette remains stable in size, no pleural effusion or pneumothorax is present,



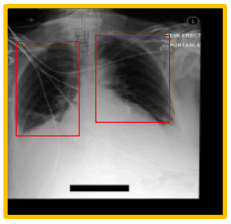
Atelectasis

Ground truth

Portable supine AP view of the chest provided demonstrates an endotracheal tube with tip positioned approximately 3,5 cm above the carina. The NG tube courses into the left upper abdomen. There is bibasilar atelectasis. Heart and mediastinal contour appears grossly unremarkable. The bony structures appear intact.

Ours

the portable supine ap chest view shows an endotracheal tube with its tip positioned approximately 3,5 cm above the carina, the ng tube extends into the left upper abdomen, bilateral atelectasis is observed, the heart and mediastinal contour seem generally unremarkable, and the bony structures are intact,



Atelectasis, enlarged heart

Ground truth

The ET tube terminates approximately 2,9 cm from the carina. The NG tube courses below the diaphragm with the tip out of the field of view of the film. There has been interval worsening of the right linear opacification likely secondary to atelectasis. No pneumothorax or definite pleural effusion is seen. The hilar and mediastinal contours are normal. There is mild cardiomegaly, stable compared to the preior exam.

Ours

the et tube ends roughly 2,9 cm away from the carina, the ng tube extends below the diaphragm, with its tip not visible on the film, there has been a gradual worsening of right sided linear opacity, possibly due to atelectasis, no signs of pneumothorax or pleural effusion are present, the hilar and mediastinal structures appear normal, mild cardiomegaly is noted, but it remains stable compared to the previous examination,

Fig. 6. The report generation results on MIMIC-CXR. Light green represents correct predictions, and red represents incorrect predictions, and the strikethrough represents the results that have not been generated.

sample distribution bias, and CTA—being a specialized angiographic description not central to MIMIC-CXR's focus on routine chest X-ray pathologies (Section 5.1, Datasets)—appears infrequently. This scarcity prevents the HMVF model (particularly its Disease Label Extractor with Elimination-Multi-Head Attention, Section 3.1) from learning sufficient associations between CTA-related visual features and text, leading to weakened weighting of CTA-corresponding image patches and subsequent failure to generate the term. Second, the misinterpretation of "periodic deterioration" arises from cross-modal semantic alignment bias: as explained in Section 3.3 (Cross-Modal Report Generator), the model relies on global visual-text feature matching, but "periodic" and "gradual" share similar visual manifestations of deterioration (e.g., lesion expansion) in chest X-rays, making it hard to distinguish their temporal semantic differences. Additionally, Section 5.2 (Implementation Details) notes that low-frequency terms (likely including "periodic deterioration") are prone to being overshadowed by more common phrases like "gradual deterioration" during training, as the model prioritizes statistically frequent text patterns. These issues collectively reflect the inherent contradictions between data scarcity, visual feature ambiguity, and fine-grained semantic alignment in medical report generation, as discussed in Section 6.1 (Limitations) of the manuscript.

5.7. Ablation Study

In this section, we further evaluated the impact of each component on the overall performance of the HMVF network and conducted tests on the IU-Xray and MIMIC-CXR datasets. Table 4 below shows the impact of each feature flow branch, where "base" only represents the HMVF network with two lesion information extractor branches. "Disease label extractor" and "Lung bounding box extractor" represent the disease label extractor branch and the bounding box locator branch respectively. Ours represents the complete HMVF framework.

As can be seen from Table 4, our model achieved the best performance on both datasets and all metrics. Specifically, the Base model produced the worst performance. The introduction of the bounding box locator enhanced the ability of HMVF to extract image features, alleviated the information extraction pressure of the lesion information analyzer, and achieved a performance improvement. Subsequently, by introducing the FFAN to adjust the visual features of different branches, it made them more suitable for the report generator, thus improving the evaluation scores again. At the very least, it can be said that our model can indeed identify the general lung organs, thereby providing lung region clues and disease label clues for the lesion information extractor and providing more reference location information when aligning with the text at the report generator stage.

Table 5 presents the model performance when using different alignment optimization methods on the IU X-Ray and MIMIC-CXR datasets. "Base" refers to using only the traditional text loss, while C-Loss and B-Loss represent the losses of the label branch and the bounding box branch respectively. In Table 5, different alignment methods contribute to improving the quality, and the three types of losses can coexist, which is helpful for model optimization. In addition, we found that the impacts of the lung bounding box loss and the disease label loss are different, and the impact of the lung bounding box loss is greater, which also proves the importance of our multi-level innovation. By combining Tables 4 and 5, the effectiveness and rationality of each component in HMVF are verified, demonstrating the overall integrity and fundamental principle of this method.

In addition, to comprehensively evaluate the role of the disease label extractor, we provided Grad-CAM diagrams under different ablation experiment conditions. As shown in Figure 7, after adding the bounding box extractor, the quality of the generated reports has also been correspondingly improved. Specifically, as shown in Figure 7, the label extractor only transmits information unidirectionally to the bounding box extractor. However, in the absence of the bounding box extractor, the attention of the label extractor will also decrease, leading to misdiagnosis in the report. For example, for the term "extubation". After introducing the bounding box extractor, due to the enhanced ability to extract visual features, the model generates reports with accurate pathological information, and the attention will be more concentrated on the effective area rather than the entire image. However, there will still be missing information, such as "CTA". The introduction of the features of the lesion area has significantly alleviated this problem, reducing misdiagnosis and generating more accurate case information. In addition, after introducing the features of the lesion area, the model has shown obvious progress in enhancing the attention to disease information. Finally, the model we designed has generally achieved the best report generation results, confirming the effectiveness and rationality of each component.

6. Limitations and discussion

6.1. Limitations

Scalability Issues: When dealing with larger and more complex datasets, the HMVF model may encounter scalability problems. Currently, the model's architecture encompasses multiple feature extraction branches, such as the Disease Label Extractor, Lung Border Extractor, and Lesion Information Extractor, along with the Feature Flow Adjustment Network and Cross-Modal Report Generator. As the size and complexity of the dataset increase, the computational load for processing these components will also rise significantly. For example, the Disease Label Extractor divides the image into patches and performs multiple convolution and attention operations. In the case of larger datasets, the number of patches and the complexity of operations on each patch will pose challenges to the model's processing speed. Moreover, the pre-training and fine-tuning processes of the model require substantial computational resources and time. If the dataset expands, the training time may become prohibitively long, and the memory requirements may exceed the capacity of ordinary computing devices. This will limit the application of the model in real-world scenarios that demand large-scale data processing.

Performance in Real-World Environments: In real-world settings, the HMVF model may face performance issues. Although it has demonstrated good results on the IU-Xray and MIMIC-CXR datasets, medical images in the real world can vary greatly. They may

Table 4
Ablation study on two benchmark datasets. The best results are marked in bold.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
IU-Xray	Base	0.124	0.121	0.190	0.082	0.153	0.208
	+LBE	0.321	0.323	0.201	0.138	0.159	0.289
	+LBE+FFAN	0.465	0.331	0.214	0.151	0.191	0.383
	Ours(LBE+FFAN+CLA)	**0.512**	**0.365**	**0.231**	**0.181**	**0.212**	**0.410**
MIMIC-CXR	Base	0.109	0.103	0.144	0.092	0.137	0.201
	+LBE	0.331	0.216	0.150	0.096	0.152	0.280
	+LBE+FFAN	0.357	0.215	0.155	0.101	0.159	0.270
	Ours(LBE+FFAN+CLA)	**0.391**	**0.239**	**0.161**	**0.115**	**0.160**	**0.293**

Table 5
Performance Comparison of Different Alignment Methods.

Dataset	Alignment	BLEU-4	METEOR	ROUGE-L	PRECISION
IU-Xray	Base	0.164	0.194	0.403	0.377
	+C-Loss	0.175	0.198	0.405	0.381
	+B-Loss	0.178	0.201	0.408	0.385
	All	**0.181**	**0.212**	**0.410**	0.388
MIMIC-CXR	Base	0.109	0.152	0.286	0.427
	+C-Loss	0.112	0.157	0.287	0.429
	+B-Loss	0.113	0.158	0.290	0.430
	All	**0.115**	**0.160**	**0.293**	0.431

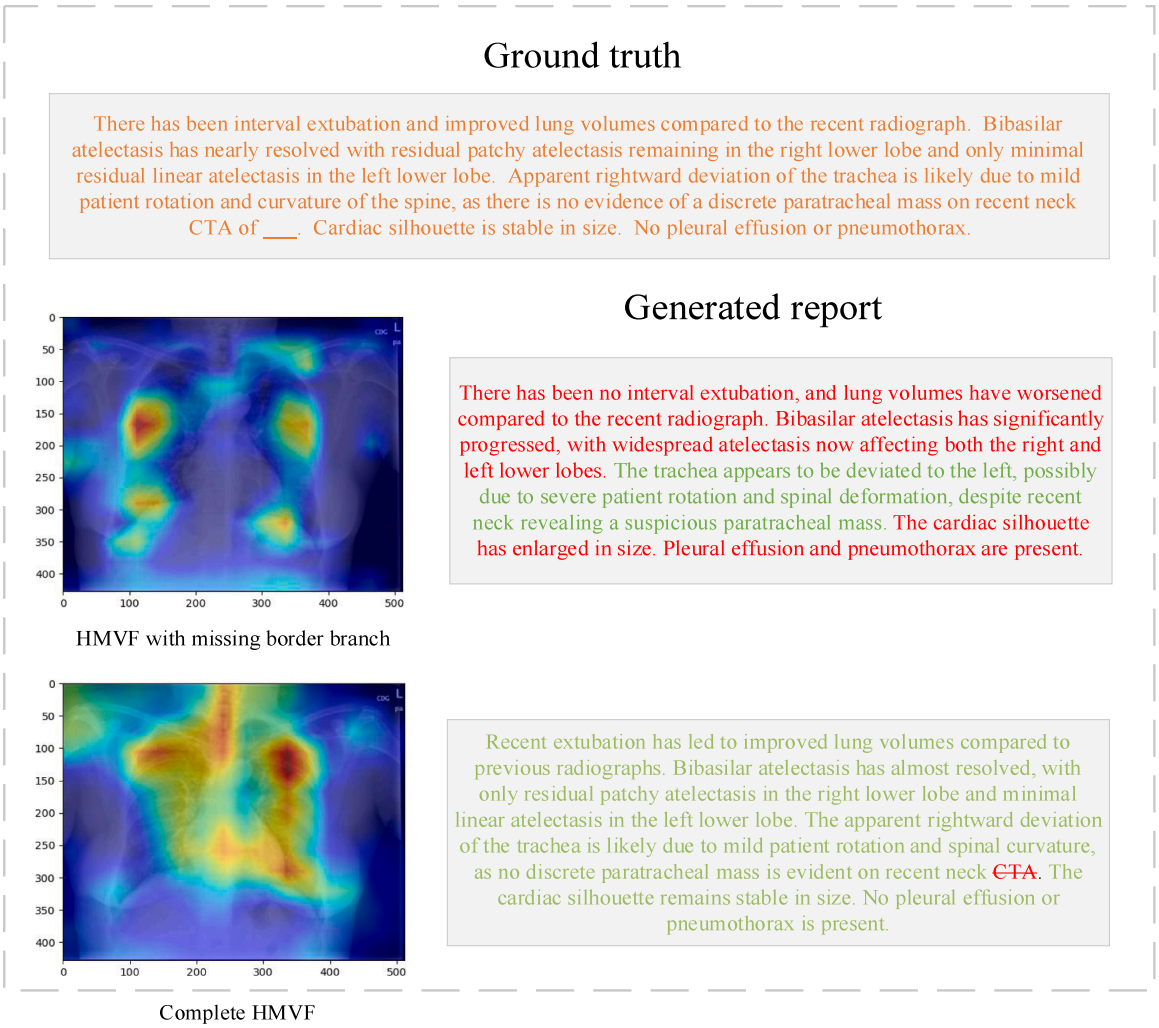


Fig. 7. Grad-CAM diagrams of the label extractor under different situations. The various representations in the figures have been described previously.

contain various artifacts, different imaging modalities, and complex disease manifestations that are not fully covered by existing datasets. For instance, in some clinical cases, medical images may be affected by patient movement during imaging, resulting in blurred images. In such situations, the model may have difficulty accurately extracting features and generating reliable reports. Additionally, the model's reliance on specific datasets means that it may not generalize well to new and unseen data distributions. If the distribution of real-world data differs from that of the training data, the performance of the model, especially in terms of accurate disease diagnosis and report generation, may decline.

Data-related Limitations: Compared with other types of datasets, medical report datasets are scarce, and this scarcity affects the HMVF model. The limited data restricts the model's ability to learn diverse medical report patterns. When handling larger and more

complex datasets, the insufficiency of data becomes more prominent. There may not be enough examples of rare diseases or complex medical conditions in the dataset, which could lead to inaccurate or incomplete reports for such cases. Furthermore, the collection of medical data is often hampered by privacy and ethical issues. This makes it difficult to expand the dataset, further limiting the model's ability to improve its performance on more complex data.

Model Complexity and Interpretability: While the complexity of the HMVF model contributes to its performance in certain tasks, it also brings challenges. The multiple interconnected components and complex operations make the model difficult to interpret. In real-world medical settings, interpretability is of great significance for doctors and medical professionals. They need to understand how the model arrives at a specific diagnosis or report. However, with the HMVF model, it is not straightforward to trace the decision-making process. For example, when the model generates a report, it is challenging to determine which specific features from which branches have the most significant impact on the final result. This lack of interpretability may reduce doctors' trust in the model, thereby limiting its practical application in clinical practice.

Challenges in Cross-Modal Alignment: Although the HMVF model attempts to address the cross-modal alignment between visual features and text embeddings, limitations still exist. In more complex scenarios, such as when dealing with images in reports that involve multiple co-existing diseases and complex medical terminologies, the model may not be able to fully align the relevant information. As can be seen from the qualitative analysis, the model sometimes misinterprets medical concepts during the report-generation process. As datasets become larger and more complex, the number of such misinterpretations may increase because the model has to handle a greater variety of visual-text relationships. This may lead to a decrease in the accuracy and reliability of medical reports, which is a concern in real-world medical applications.

6.2. Discussion

Although some progress has been made in the current automatic generation technology of medical reports, there are still some problems in this field. For example, compared with image-classification tasks, medical report datasets are scarce, and there are many obstacles to creating medical imaging report datasets, involving privacy and ethical issues. Secondly, the collectable data samples are limited. Even if cases are collected from multiple hospitals or institutions, although there are general medical report writing formats to follow, due to the free-style writing characteristics of medical report texts and the subjectivity of diagnosing doctors, the styles of medical reports will vary greatly, thus increasing the difficulty of text report formatting and cross-modal alignment in the model training process. In addition, the number of negative samples(healthy images) in medical imaging reports is much larger than that of positive samples (disease images), which also brings certain challenges to disease analysis. In future work, we will also try to build our own medical report dataset and explore the possibilities of more medical report generation methods.

7. Conclusion

In this paper, we propose an effective medical report-generation method, HMVF. This method extracts visual features from multiple hierarchical branches from different angles, and achieves high-quality medical report generation through cross-modal alignment after information enhancement.

For most medical report-generation tasks, networks based on the Transformer architecture or the diagnostic processes of doctors or professional teams are used to generate medical imaging reports. Based on this, we break away from the diagnostic processes of doctors or medical teams and innovatively introduce hierarchical feature representations to capture organ lesion information, thereby enhancing the model's understanding of medical images. At the same time, we transmit visual data streams containing different information features to different layers of the report generator and align them with text embeddings simultaneously, which increases the model's understanding of medical images and cross-modal alignment ability. The experimental results show that, compared with existing methods, our method exhibits excellent performance in comprehensively generating realistic and reliable medical reports.

In future work, we will attempt to build our own medical report dataset and explore more feasible medical report-generation methods that can improve NLG (Natural Language Generation) and accuracy scores, jointly promoting the development of the medical report-generation field.

Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (No. 61966026) and the Natural Science Foundation of Inner Mongolia Autonomous Region (No.2020MS06015).

Data availability

https://openi.nlm.nih.gov/
https://physionet.org/content/mimic-cxr-jpg/2.0.0/

CRediT authorship contribution statement

Hailong Zuo: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Zhi Weng:** Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition,

Formal analysis, Data curation, Conceptualization. **Yunlu Duan:** Validation, Supervision, Formal analysis. **Zijing Lv:** Validation, Formal analysis. **Feifan Bi:** Supervision, Formal analysis, Data curation. **Zhiqiang Zheng:** Funding acquisition.

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.

References

[1] Z. Gao, J. Guo, L. Chen, K. Wang, Y. Chen, Y. Ke, S. Yang, AnDR-BLIP2: enhanced semantic understanding framework for industrial image anomaly detection and report generation, Journal of the Franklin Institute 362 (2025) 107816, https://doi.org/10.1016/j.jfranklin.2025.107816.

[2] H. Zhuang, Y. Yan, R. He, Z. Zeng, Class incremental learning with analytic learning for hyperspectral image classification, Journal of the Franklin Institute 361 (2024) 107285, https://doi.org/10.1016/j.jfranklin.2024.107285.

[3] K. Han, M. Chen, C. Gao, C. Qing, DCA-Unet: Enhancing small object segmentation in hyperspectral images with Dual Channel Attention Unet, Journal of the Franklin Institute 362 (2025) 107532, https://doi.org/10.1016/j.jfranklin.2025.107532.

[4] H. Wang, J. Wang, Z. Fan, Deep convolutional sparse dictionary learning for bearing fault diagnosis under variable speed condition, Journal of the Franklin Institute 362 (2025) 107392, https://doi.org/10.1016/j.jfranklin.2024.107392.

[5] Z. Zhang, Q. Yang, S. He, J. Chen, Multi-level semantic-aware communication for multi-task image transmission, Journal of the Franklin Institute 362 (2025) 107598, https://doi.org/10.1016/j.jfranklin.2025.107598.

[6] H. Pan, X. Li, H. Ge, L. Wang, X. Yu, Multi-scale hierarchical cross fusion network for hyperspectral image and LiDAR classification, Journal of the Franklin Institute 362 (2025) 107713, https://doi.org/10.1016/j.jfranklin.2025.107713.

[7] L. Min, Z. Zhang, Z. Jin, Selective segmentation of inhomogeneous images based on local clustering and global smoothness, Journal of the Franklin Institute 362 (2025) 107591, https://doi.org/10.1016/j.jfranklin.2025.107591.

[8] H. Pan, H. Yan, H. Ge, M. Liu, C. Shi, Transformer-enhanced two-stream complementary convolutional neural network for hyperspectral image classification, Journal of the Franklin Institute 361 (2024) 106973, https://doi.org/10.1016/j.jfranklin.2024.106973.

[9] D. Gao, M. Kong, Y. Zhao, J. Huang, Z. Huang, K. Kuang, F. Wu, Q. Zhu, Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning, Medical Image Analysis 91 (2024) 102982, https://doi.org/10.1016/j.media.2023.102982.

[10] T. Tanida, P. Müller, G. Kaissis, D. Rueckert, Interactive and Explainable Region-guided Radiology Report Generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7433–7442, https://doi.org/10.1109/CVPR52729.2023.00718.

[11] Q. Tang, L. Xu, Y. Wang, B. Zheng, J. Lv, X. Zeng, W. Li, Dual-modality visual feature flow for medical report generation, Medical Image Analysis 101 (2025) 103413, https://doi.org/10.1016/j.media.2024.103413.

[12] S. Yang, X. Wu, S. Ge, S.K. Zhou, L. Xiao, Knowledge matters: Chest radiology report generation with general and specific knowledge, Medical Image Analysis 80 (2022) 102510, https://doi.org/10.1016/j.media.2022.102510.

[13] A. Liu, Y. Guo, J.-H. Yong, F. Xu, Multi-Grained Radiology Report Generation With Sentence-Level Image-Language Contrastive Learning, IEEE Transactions on Medical Imaging 43 (2024) 2657–2669, https://doi.org/10.1109/TMI.2024.3372638.

[14] O. Akinlade, E. Vakaj, A. Dridi, S. Tiwari, F. Ortiz-Rodriguez, Semantic Segmentation of the Lung to Examine the Effect of COVID-19 Using UNET Model, in: M. A. Jabbar, F. Ortiz-Rodríguez, S. Tiwari, P. Siarry (Eds.), Applied Machine Learning and Data Analytics, Springer Nature Switzerland, 2023, pp. 52–63, https://doi.org/10.1007/978-3-031-34222-6_5.

[15] G. Taiwo, S. Vadera, A. Alameer, Vision transformers for automated detection of pig interactions in groups, Smart Agricultural Technology 10 (2025) 100774, https://doi.org/10.1016/j.atech.2025.100774.

[16] Y. Yang, J. Yu, Z. Fu, K. Zhang, T. Yu, X. Wang, H. Jiang, J. Lv, Q. Huang, W. Han, Token-Mixer: Bind Image and Text in One Embedding Space for Medical Image Reporting, IEEE Transactions on Medical Imaging 43 (2024) 4017–4028, https://doi.org/10.1109/TMI.2024.3412402.

[17] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach, IEEE Transactions on Medical Imaging 41 (2022) 2803–2813, https://doi.org/10.1109/TMI.2022.3171661.

[18] X. Mei, L. Yang, D. Gao, X. Cai, J. Han, T. Liu, PhraseAug: An Augmented Medical Report Generation Model With Phrasebook, IEEE Transactions on Medical Imaging 43 (2024) 4211–4223, https://doi.org/10.1109/TMI.2024.3416190.

[19] S. Bu, T. Li, Y. Yang, Z. Dai, Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14194–14204, https://doi.org/10.1109/CVPR52733.2024.01346.

[20] S. Yang, X. Wu, S. Ge, Z. Zheng, S.K. Zhou, L. Xiao, Radiology report generation with a learned knowledge base and multi-modal alignment, Medical Image Analysis 86 (2023) 102798, https://doi.org/10.1016/j.media.2023.102798.

[21] I. Najdenkoska, X. Zhen, M. Worring, L. Shao, Uncertainty-aware report generation for chest X-rays by variational topic inference, Medical Image Analysis 82 (2022) 102603, https://doi.org/10.1016/j.media.2022.102603.

[22] S. Yan, W.K. Cheung, K. Chiu, T.M. Tong, K.C. Cheung, S. See, Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation, IEEE Transactions on Medical Imaging 42 (2023) 2211–2222, https://doi.org/10.1109/TMI.2023.3245608.

[23] Z. Huang, X. Zhang, S. Zhang, KiUT: Knowledge-injected U-Transformer for Radiology Report Generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19809–19818, https://doi.org/10.1109/CVPR52729.2023.01897.

[24] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, X. Chang, Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3334–3343, https://doi.org/10.1109/CVPR52729.2023.00325.

[25] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13748–13757, https://doi.org/10.1109/CVPR46437.2021.01354.

[26] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310, https://doi.org/10.1093/jamia/ocv080.

[27] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C. Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, arXiv E-Prints (2019) arXiv:1901.07042. 10.48550/arXiv.1901.07042.

[28] W. Indeewara, M. Hennayake, K. Rathnayake, T. Ambegoda, D. Meedeniya, Chest X-ray Dataset with Lung Segmentation, (n.d.). 10.13026/9CY4-F535.

[29] Z. Chen, Y. Shen, Y. Song, X. Wan, 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 1, Association for Computational Linguistics, 2021, pp. 5904–5914, https://doi.org/10.18653/v1/2021.acl-long.459. Long PapersOnline.

[30] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, Philadelphia, Pennsylvania, Association for Computational Linguistics, 2001, p. 311, https://doi.org/10.3115/1073083.1073135.

[31] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: IEEvaluation@ACL, 2005. https://api.semanticscholar.org/CorpusID:7164502.

[32] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Annual Meeting of the Association for Computational Linguistics, 2004. https://api.semanticscholar.org/CorpusID:964287.
[33] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical Sequence Training for Image Captioning, arXiv E-Prints (2016) arXiv:1612.00563. 10.48550/arXiv.1612.00563.
[34] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning, arXiv E-Prints (2016) arXiv:1612.01887. 10.48550/arXiv.1612.01887.
[35] Y. Li, X. Liang, Z. Hu, E.P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, Advances in Neural Information Processing Systems 31 (2018).
[36] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating Radiology Reports via Memory-driven Transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 1439–1449, https://doi.org/10.18653/v1/2020.emnlp-main.112. Online.
[37] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164, https://doi.org/10.1109/CVPR.2015.7298935.
[38] B. Jing, Z. Wang, E. Xing, 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, Florence, Italy, Association for Computational Linguistics, 2019, pp. 6570–6580, https://doi.org/10.18653/v1/P19-1657.
[39] B. Jing, P. Xie, E. Xing, On the Automatic Generation of Medical Imaging Reports, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia 1, Association for Computational Linguistics, 2018, pp. 2577–2586, https://doi.org/10.18653/v1/P18-1240. Long Papers).
[40] Y. Wang, Z. Lin, Z. Xu, H. Dong, J. Tian, J. Luo, Z. Shi, Y. Zhang, J. Fan, Z. He, Trust It or Not: Confidence-Guided Automatic Radiology Report Generation, arXiv E-Prints (2021) arXiv:2106.10887. 10.48550/arXiv.2106.10887.
[41] J. Wang, A. Bhalerao, T. Yin, S. See, Y. He, CAMANet: Class Activation Map Guided Attention Network for Radiology Report Generation, arXiv E-Prints (2022) arXiv:2211.01412. 10.48550/arXiv.2211.01412.
[42] P. Divya, Y. Sravani, C. Vishnu, C.K. Mohan, Y.W. Chen, Memory Guided Transformer With Spatio-Semantic Visual Extractor for Medical Report Generation, IEEE Journal of Biomedical and Health Informatics 28 (2024) 3079–3089, https://doi.org/10.1109/JBHI.2024.3371894.
[43] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: 2017: pp. 375–383.
[44] C.Y. Li, X. Liang, Z. Hu, E.P. Xing, Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation, arXiv E-Prints (2018) arXiv:1805.08298. 10.48550/arXiv.1805.08298.
[45] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation, arXiv E-Prints (2021) arXiv:2106.06963. 10.48550/arXiv.2106.06963.
[46] Y. Xue, T. Xu, L.Rodney Long, Z. Xue, S. Antani, G.R. Thoma, X. Huang, Multimodal Recurrent Model with Attention for Automated Radiology Report Generation, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, 2018, pp. 457–466.
[47] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating Radiology Reports via Memory-driven Transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 1439–1449, https://doi.org/10.18653/v1/2020.emnlp-main.112. Online.
[48] Z. Chen, Y. Shen, Y. Song, X. Wan, 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 1, Association for Computational Linguistics, 2021, pp. 5904–5914, https://doi.org/10.18653/v1/2021.acl-long.459. Long Papers)Online.
[49] H. Qin, Y. Song, Reinforced Cross-modal Alignment for Radiology Report Generation. Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, 2022, pp. 448–458, https://doi.org/10.18653/v1/2022.findings-acl.38.
[50] S. Yang, X. Wu, S. Ge, S.K. Zhou, L. Xiao, Radiology Report Generation with a Learned Knowledge Base and Multi-modal Alignment, (2021). 10.48550/ARXIV.2112.15011.