APSIPA Transactions on Signal and Information Processing, 2024, 13, e24

Original Paper

Automatic Medical Report Generation: Methods and Applications

Li Guo*, Anas M. Tahir, Dong Zhang, Z. Jane Wang and Rabab K. Ward

Electrical and Computer Engineering Department, University of British Columbia, Vancouver, Canada

ABSTRACT

The increasing demand for medical imaging has surpassed the capacity of available radiologists, leading to diagnostic delays and potential misdiagnoses. Artificial intelligence (AI) techniques, particularly in automatic medical report generation (AMRG), offer a promising solution to this dilemma. This review comprehensively examines AMRG methods from 2021 to 2024. It (i) presents solutions to primary challenges in this field, (ii) explores AMRG applications across various imaging modalities, (iii) introduces publicly available datasets, (iv) outlines evaluation metrics, (v) identifies techniques that significantly enhance model performance, and (vi) discusses unresolved issues and potential future research directions. This paper aims to provide a comprehensive understanding of the existing literature and inspire valuable future research.

Keywords: Medical report generation, deep learning, artificial intelligence, review

1 Introduction

Automatic medical report generation (AMRG) is an emerging research area in artificial intelligence (AI) within the medical field [99, 55]. It utilizes computer

*Corresponding author: Li Guo, lguo@ece.ubc.ca.

Received 11 July 2024; revised 16 August 2024; accepted 02 September 2024
ISSN 2048-7703; DOI 10.1561/116.20240044

vision (CV) and natural language processing (NLP) to interpret medical images and generate descriptive, human-like reports. AMRG has been applied to various imaging modalities, including X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI), and ultrasound [47, 7, 2]. This technology has the potential to streamline the diagnostic process, alleviate the workload on radiologists, and enhance diagnostic accuracy.

Traditionally, the interpretation of medical images relies on trained radiologists, a labor-intensive and error-prone process [14, 5, 138, 4]. In the US and UK, the number of radiologists is insufficient to meet the growing demand for imaging and diagnostics [106, 105]. In resource-poor regions, the scarcity of radiology services is even more severe [48, 107]. This shortage of radiologists leads to delays and backlogs in interpreting medical images. In 2015, approximately 330,000 patients in the UK waited more than 30 days for radiology reports [88]. Due to delayed reports, some urgent images have to be reviewed by emergency physicians. However, the discernible interpretation differences between emergency physicians and trained radiologists can lead to missed diagnoses and misdiagnoses [36]. Additionally, reports written by professional radiologists exhibit a 3-5% error rate and approximately 35% uncertainty rate [12, 114, 79]. As workloads increase, the probability of errors by radiologists also rises [33, 62]. For instance, an American doctor was sued after failing to detect a case of breast cancer due to reading too many X-rays in one day [11]. AMRG addresses these issues by providing a systematic approach to image interpretation, potentially improving diagnostic efficiency and accuracy.

In recent years, deep learning has made significant progress in image analysis, with convolutional neural networks (CNNs) and Transformers excelling in high-precision lesion detection and classification of medical conditions [115, 39, 64]. NLP techniques translate visual information from medical images into natural language reports, covering imaging findings, diagnostic conclusions, and recommendations, thereby achieving seamless image-to-text conversion [17, 111, 35]. Researchers have developed various AMRG methods by combining CNNs, Transformers, and NLP in an encoder-decoder architecture [99, 3, 10].

Despite these advancements, this field still faces numerous challenges. Firstly, bridging the modal gap between image input and text output is a fundamental challenge for AMRG. Medical images contain complex information that must be accurately interpreted and translated into coherent text, requiring sophisticated algorithms to map visual patterns to medical terminology. Secondly, medical images exhibit unique visual deviations: lesion areas usually occupy a small portion of the image, leading to highly similar normal and abnormal images, necessitating AMRG systems to be more sensitive to fine-grained differences than general image captioning models [157, 13, 37]. Thirdly, medical reports are long texts with high clinical professionalism and accuracy, placing higher standards on the quality of the generated texts, which

demands advanced NLP techniques to handle detailed and precise medical documentation [58, 1, 31]. Finally, medical datasets are limited and noisy; datasets like MIMIC-CXR (0.22M) [56] and IU-Xray (4k) [27] are smaller than image recognition datasets like ImageNet (14M) [28] and image captioning datasets like Conceptual Captions (3.3M) [110], limiting model training effectiveness. Furthermore, noise in medical reports, such as temporal information, can confuse models and lead to inaccuracies or hallucinations [104, 9].

In this review, we comprehensively examine 112 papers on AMRG based, predominantly from 2021 to 2024, and summarize various solutions proposed to address the aforementioned challenges. Our scope extends beyond radiographic report generation to include emerging applications in modalities such as MRI, CT, and ultrasound. Additionally, we present the public datasets and evaluation metrics used in this field. Through a comparative analysis of state-of-the-art (SOTA) models on benchmark radiography datasets, we identify techniques that significantly enhance evaluation metrics. Finally, we discuss potential future directions for this field. The structure of this paper is illustrated in Figure 1.

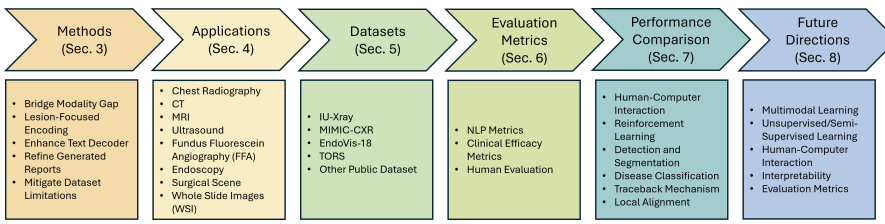

Figure 1: The content road map of this review paper. First, we present five types of solutions to address the challenges of AMRG. Next, we explore the applications of AMRG across different imaging modalities. Following this, we introduce various public datasets. Then, we outline the evaluation metrics employed to assess model performance. By comparing the performance of models on benchmark datasets, we identify six techniques that effectively enhance model performance. Finally, we discuss future research directions in the field.

2 Problem Statement

The objective of AMRG is to train a model that can extract meaningful features from medical images and generate descriptive text sequences that accurately describe the medical conditions depicted in the images. The primary objective function is the word-level cross-entropy loss, which measures the discrepancy between the predicted word probabilities and the actual words in the ground truth (GT) reports.

Figure 2 illustrates the basic structure of the AMRG model. Given a medical image, the image encoder extracts a sequence of image features I. The

text decoder, which can be either an RNN or a Transformer model, generates a sequence of words $\{w_1, w_2, ..., w_T\}$ to describe the medical image in an autoregressive manner. At each time step t, the decoder generates the next word w_t based on the previous words $\{w_1, w_2, ..., w_{t-1}\}$ and image features I. Assuming that the GT report is $\{w_1^*, w_2^*, ..., w_T^*\}$, the cross-entropy loss at each time step t is given by:

$$\mathcal{L}_{CE}(t) = -\log P(w_t^* | w_1^*, ..., w_{t-1}^*, I) \tag{1}$$

The total loss for the entire sequence is the sum of the losses over all time steps:

$$\mathcal{L}_{CE} = \sum_{t=1}^{T} \mathcal{L}_{CE}(t) = -\sum_{t=1}^{T} \log P(w_t^* | w_1^*, ..., w_{t-1}^*, I) \tag{2}$$

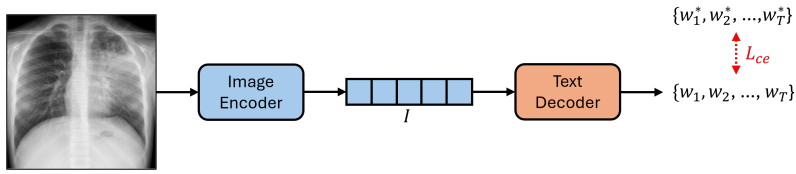

Figure 2: Schematic of the basic report generation model. An image encoder extracts features I from the input image, which are then processed by a text decoder to generate the predicted report $\{w_1, w_2, ..., w_T\}$. The model is optimized using cross-entropy loss $\mathcal{L}_{CE}$ between the predicted and ground truth report $\{w_1^*, w_2^*, ..., w_T^*\}$.

3 Methods

In this section, we introduce various methods designed to address the aforementioned challenges. First, we discuss techniques for bridging the gap between image-text modalities (Section 3.1). Next, we present lesion-focused image encoding methods that enhance the model's ability to detect and emphasize clinically significant regions (Section 3.2). We then detail approaches for enhancing the text decoder with additional information (Section 3.3) and refining generated reports to ensure high-quality medical outputs (Section 3.4). Finally, we cover strategies to mitigate dataset flaws, including methods to handle noisy and limited datasets (Section 3.5). The four main challenges and their corresponding solutions are shown in Figure 3. The following subsections delve into these solutions in detail.

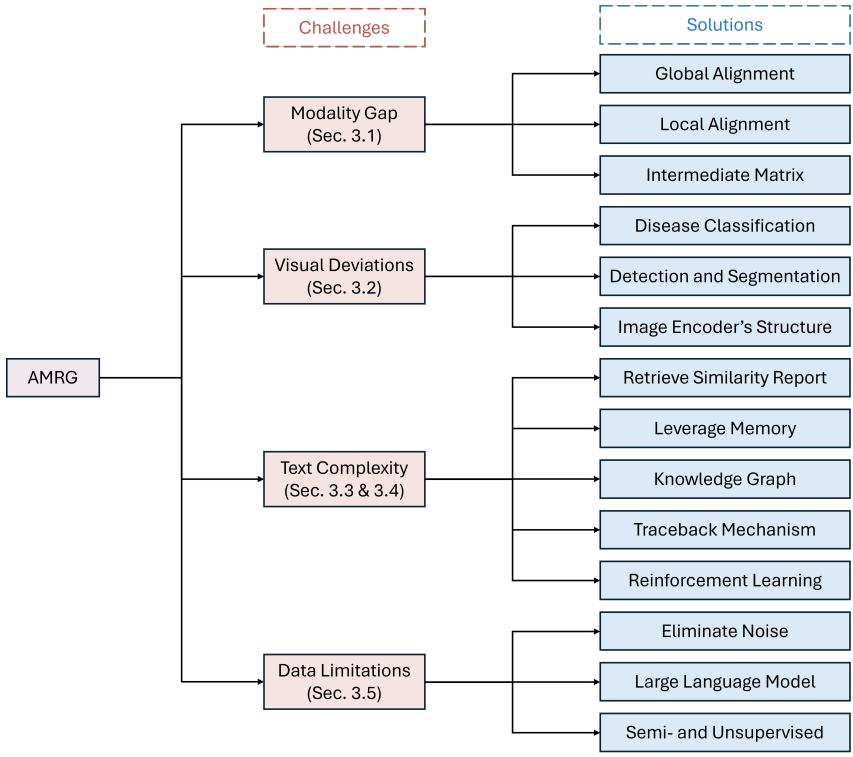

Figure 3: Four challenges in automatic medical report generation (AMRG) and their corresponding solutions.

3.1 Bridging the Gap Between Modalities

Bridging the gap between image and text modalities is crucial for medical report generation. This section introduces three key methods to address this challenge: (i) global alignment (Section 3.1.1), (ii) local alignment (Section 3.1.2), and (iii) intermediate matrix alignment (Section 3.1.3). Each method offers a distinct strategy for aligning visual and textual data. Global alignment focuses on aligning entire images with entire reports to maximize mutual information and minimize discrepancies. Local alignment targets fine-grained interactions by associating specific image regions with textual elements such as sentences or words. Intermediate matrix alignment employs a shared learnable matrix to capture the alignment between visual and textual features. Figure 4 presents simplified flowcharts of these three alignment methods. The following subsections provide detailed explanations of each method.

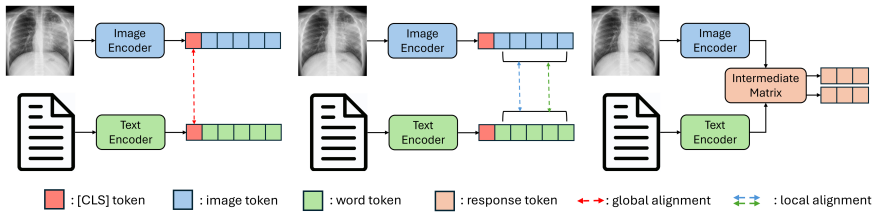

Figure 4: Flowcharts of three representative alignment methods. The left diagram illustrates global alignment, which typically uses the [CLS] token to represent the global representation of a modality. The middle diagram depicts local alignment, aligning image patches with word tokens. The right diagram shows alignment via an intermediate matrix, where a shared matrix represents the features of both modalities, ensuring they are in the same latent space.

3.1.1 Global Alignment

Global alignment is a method that aligns the entire image with the entire report based on InfoNCE loss [96] and triplet loss [108]. InfoNCE loss is well-suited for large datasets because it processes all negative pairs in a batch. In contrast, triplet loss specializes in identifying fine-grained differences by focusing on individual negative samples at a time.

The infoNCE loss function creates a joint embedding space by maximizing the cosine similarity between positive image-text pairs and minimizing it for negative pairs. This process closely aligns images and their corresponding reports. The CLIP framework [103], which pioneers the use of InfoNCE loss for visual representation learning under natural language supervision, is particularly beneficial for report generation. This approach ensures that each medical image is effectively supervised by its paired report. Several studies have employed CLIP loss (InfoNCE loss) to successfully mitigate the modality discrepancies between radiographic images and clinical reports [154, 9, 30, 86, 158]. Specifically, given a batch of N pairs of image embeddings $\{I_i\}$ and text embeddings $\{T_i\}$, the CLIP loss can be formulated as follows:

$$\begin{aligned}
\mathcal{L}_{IN}^{I\to T}(I,T) &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S(I_i,T_i)/\tau_1)}{\sum_{j=1}^{N}\exp(S(I_i,T_j)/\tau_1)} \\
\mathcal{L}_{IN}^{T\to I}(I,T) &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S(T_i,I_i)/\tau_1)}{\sum_{j=1}^{N}\exp(S(T_i,I_j)/\tau_1)} \\
\mathcal{L}_{CL}(I,T) &= \frac{1}{2}(\mathcal{L}_{IN}^{I\to T}(I,T)+\mathcal{L}_{IN}^{T\to I}(I,T)),
\end{aligned} \tag{3}$$

where τ_1 is a temperature parameter that scales the logits and $S(\cdot,\cdot)$ denotes cosine similarity.

However, CLIP's single-view supervision inadequately captures the intricate semantic relationships between images and text. To address this limitation, CXR-CLIP [149] employs a multi-view supervision (MVS) technique [74] that enhances training efficacy by incorporating multiple views. For instance, each chest X-ray study includes images from both the postero-anterior and lateral views (denoted as I^1, I^2), along with two text descriptions, findings, and impressions (denoted as T^1, T^2). The CLIP loss is expanded to MVS loss:

$$\mathcal{L}_{MVS} = \frac{1}{4}(\mathcal{L}_{CL}(I^1, T^1) + \mathcal{L}_{CL}(I^2, T^1) + \mathcal{L}_{CL}(I^1, T^2) + \mathcal{L}_{CL}(I^2, T^2)). \quad (4)$$

Furthermore, the triplet loss, another significant contrastive loss function, ensures that an anchor sample's embedding is closer to a positive sample than any negative sample by at least a predefined margin α. This loss function has been particularly effective in the medical field, enhancing discrimination between closely resembling reports and images [131, 75]. Based on the paired image embeddings I and text embeddings T extracted by two unimodal encoders, the hardest negative samples $\tilde{I}$, $\tilde{T}$ in the batch of N pairs are selected by their highest similarity to the corresponding GT modality. The triplet loss is optimized as follows:

$$\mathcal{L}_{triplet} = \frac{1}{N}\sum_{i=1}^{N}[\alpha - S(I,T) + S(I,\tilde{T})]_+ + [\alpha - S(I,T) + S(\tilde{I},T)]_+, \quad (5)$$

where α is the margin value and $[\cdot]_+$ represents the positive part (i.e., $\max(0, \cdot)$). It is noteworthy that combining InfoNCE loss and triplet loss can yield synergistic effects, enhancing model performance [53].

Moreover, recognizing the limitations of using two unimodal encoders for distinguishing hard negative samples, Li *et al.* [69] recommend a multimodal encoder strategy to explore more complex modal interactions. Specifically, the image and text embeddings are jointly input into a multimodal encoder to predict whether the image and text match. Further studies have validated this approach, confirming the robustness of the multimodal encoder strategy in medical report generation [72, 52]. Additionally, exploring the use of a text decoder to generate image-text matching scores offers another avenue for optimizing the overall model beyond just the encoders [127].

3.1.2 Local Alignment

Although global alignment is an effective and widely adopted method, contrasting the entire image with the entire report can result in overlooking fine-grained interactions between different modalities. To address this limitation, researchers have introduced two local alignment strategies: sentence-region alignment and word-region alignment.

Sentence-region alignment matches specific image regions to corresponding sentences within a report. PhenotypeCLIP [125] employs cross-attention to generate sentence-based local textual and visual representations, replacing the global representations used in the InfoNCE loss (Equation 3) to enhance contrastive learning. Further refining this approach, PRIOR [24] replaces the softmax function in the cross-attention mechanism with a sigmoid function, which generates a sparser matrix and enhances computational efficiency. Moreover, PRIOR substitutes the InfoNCE loss, traditionally used in report-to-image local alignment, with a loss function based on cosine similarity and asymmetrical projection. This modification mitigates the risk of feature collapse, which results from the misclassification of positive image regions as negative. Specifically, given a batch of N pairs of image embeddings $I = \{I_1, I_2, ..., I_N\}$ and report embeddings $T = \{T_1, T_2, ..., T_N\}$, each image embedding I_i is composed of patches $I_i = \{I_i^1, I_i^2, ..., I_i^V\}$, and each report embedding T_i is composed of sentences $T_i = \{T_i^1, T_i^2, ..., T_i^U\}$. Here, V and U represent the number of image patches and sentences within a report, respectively. For each sentence u, the attention-based visual representation is formulated as:

$$c_i^u = \sum_{v=1}^{V} \sigma(\frac{Q^I T_i^u \cdot K^I I_i^v}{\sqrt{D}}) V^I I_i^v. \tag{6}$$

Similarly, for each image region v, the attention-based textual representation is formulated as:

$$c_i^v = \sum_{u=1}^{U} \sigma(\frac{Q^R I_i^v \cdot K^R T_i^u}{\sqrt{D}}) V^R T_i^u, \tag{7}$$

where Q^I, K^I, V^I, Q^R, K^R, V^R are learnable matrices, $\sigma(\cdot)$ is the sigmoid function, and D is the dimension of embeddings. The new report-to-image local alignment loss is formulated as:

$$\mathcal{L}_l^{T \to I} = -\frac{1}{NV} \sum_{i=1}^{N} \sum_{v=1}^{V} \frac{1}{2} [S(h(I_i^v), SG(c_i^v)) + S(h(c_i^v), SG(I_i^v))], \tag{8}$$

where h is a MLP head and SG denotes the stop-gradient operation.

Considering that image embeddings $I_i = \{I_i^1, I_i^2, ..., I_i^V\}$ and report embeddings $T_i = \{T_i^1, T_i^2, ..., T_i^U\}$ still exhibit significant differences, Liu *et al.* [80] introduce intermediate topics (anatomical entities) to further encode I_i and T_i into topic features $I_i^* = \{I_i^{*1}, I_i^{*2}, ..., I_i^{*M}\}$ and $T_i^* = \{T_i^{*1}, T_i^{*2}, ..., T_i^{*M}\}$, where M is the number of topics, and use weighted summation to obtain the local visual representation c_i, which is then utilized for local alignment:

$$I_i^* = Transformer(I_i),\ T_i^* = softmax(l(T_i)) \in \mathbb{R}^M,\ c_i = \sum_{m=1}^{M} T_i^{*m} I_i^{*m}, \tag{9}$$

where l is the linear projection.

Besides sentences, words also possess varying significance within a report. For example, descriptions of abnormalities are more important than descriptions of normal findings. Therefore, researchers have proposed word-region alignment to capture more fine-grained multimodal interactions [45, 9, 26, 20]. GLoRIA [45] learns attention weights that prioritize different image regions based on their relevance to a given word and implements local contrastive learning using attention-weighted image representations. Specifically, given a pair of image embeddings $I = \{I_1, I_2, ..., I_V\}$ and word embeddings $W = \{W_1, W_2, ..., W_T\}$, where V and T represent the number of image patches and words in a report, respectively. First, the word-region similarity s is calculated using the dot product, then softmax normalization is applied to obtain the attention weight a_{tv}. The attention-weighted sum then forms the image representation c_t for the word W_t:

$$s = I^T W, \quad a_{tv} = \frac{\exp(s_{tv}/\tau_2)}{\sum_{k=1}^{V} \exp(s_{tk}/\tau_2)}, \quad c_t = \sum_{v=1}^{V} a_{tv} I_v, \tag{10}$$

where s_{tv} denotes the similarity between the word W_t and image patch I_v, and τ_2 is a temperature parameter. Next, an aggregation function Z combines the similarities between all words W_t and their corresponding weighted image representations c_t, replacing the cosine similarity in the InfoNCE loss (Equation 3):

$$Z(I, W) = \log(\sum_{t=1}^{T} \exp(S(c_t, W_t)/\tau_3))^{\tau_3}, \tag{11}$$

where τ_3 is a temperature parameter. Dawidowicz *et al.* [26] modifies this method by substituting the dot product in Equation 10 with element-wise multiplication and using a self-attention weighted sum to aggregate the similarities instead of a simple summation in Equation 11. This modification achieves better results in various downstream tasks compared to GLoRIA.

The aforementioned methods rely on a pre-defined patch size across images. In medical image, lesions can exhibit a wide range of shapes and sizes. A fixed partition of image patches may result in incomplete or ambiguous representations of the key imaging abnormalities. Chen *et al.* [20] propose a method that splits an image into adaptive patches of variable sizes and aligns them with words. This method uses additional Transformer blocks and fully connected layers to predict the offset and new patch size for each adaptive patch. Then, it uniformly resamples feature points within these adaptive patches from the input image.

3.1.3 Intermediate Matrix

In addition to contrastive learning, another method to bridge the modal gap is the use of a learnable shared matrix to capture the alignment between images and texts. Chen *et al.* [22] propose an approach that maps visual features and textual features into a unified intermediate space. Specifically, given an embedding $I = \{I_1, I_2, ..., I_V\}$ extracted from the image and an embedding $W = \{W_1, W_2, ..., W_{t-1}\}$ extracted from the generated report, these embeddings are mapped to visual memory responses $R^I = \{R_1, R_2, ..., R_V\}$ and textual memory responses $R^W = \{R_1, R_2, ..., R_{t-1}\}$. Both R^I and R^W are derived from a shared memory matrix M. Subsequently, R^I and R^W are fed into the text decoder to generate the next word at the time step t. The effectiveness of this method has been verified by studies from Qin *et al.* [102] and You *et al.* [148].

Wang *et al.* [121] further improve this mapping method. They concatenate the embeddings of image-report pairs with the same disease label and apply K-means clustering to initialize the memory matrix. Moreover, they integrate both the image and text embeddings I and W, along with visual and textual responses R^I and R^W, and feed them into the decoder to enrich the generated content. Additionally, they incorporate triplet contrastive loss (Equation 5) into the optimization process to enhance the alignment of the visual and textual memory responses via explicit supervision signals. Similarly, Li et al. [75] also employ triplet loss to align visual and textual features post-mapping and utilize a dual-gate mechanism to more intricately fuse visual and textual features both before and after mapping.

3.2 Lesion-Focused Image Encoding

This section outlines three methods for enhancing the image encoder to focus on lesion areas and generate discriminative image representations: (i) using a classification task for joint learning (Section 3.2.1), (ii) employing pre-trained detection and segmentation networks as auxiliaries (Section 3.2.2), and (iii) modifying the internal structure of the image encoder (Section 3.2.3). The simplified flowcharts of these three methods are shown in Figure 5. The following subsections detail these approaches, elucidating how each method refines the encoder's capacity to identify and concentrate on clinically significant regions.

3.2.1 Disease Classification

Adopting the features extracted by the image encoder for multi-label disease classification is an effective joint learning strategy to adapt the encoder for the report generation task [147, 44, 141, 86, 55, 130, 123, 127, 153, 133]. This

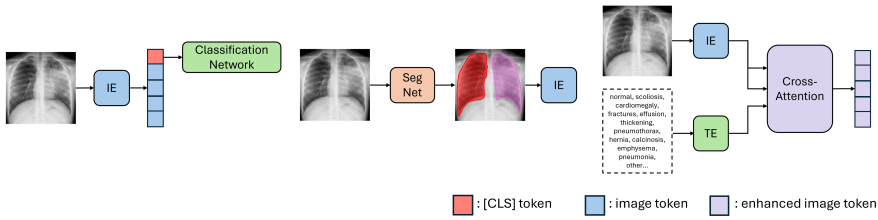

Figure 5: Flowcharts of three representative methods for enhancing image encoding: The left diagram shows that the image features extracted by the image encoder (IE) are used for disease classification, typically using only the [CLS] token instead of all image tokens. The middle diagram illustrates that the image is first processed through a pre-trained segmentation network (Seg Net) to segment meaningful areas (such as the left and right lungs), and only these areas are then input into the image encoder to eliminate background interference. The right diagram demonstrates that with cross-attention, the image features are used as keys and values, while the disease tags are used as queries. This method encourages the model to focus on image areas related to disease tags. TE represents text encoder.

strategy enables the image encoder to focus on regions where diseases are likely to occur and refine its ability to extract discriminative features, which helps decode accurate text. As a result, it enhances the encoder's sensitivity to medically relevant areas and clinically significant details indicative of various diseases.

Nevertheless, due to the distinct operational mechanisms of CNNs and Transformers, their implementation approaches exhibit slight variations. For CNN-based image encoders, all features from the last convolutional layer are used for classification [147, 44, 55], often coupled with average pooling to achieve a global representation [141, 86]. This approach may result in image features that contain only high-level information for classification, while losing the low-level information necessary for generating descriptive text. Conversely, Transformer-based image encoders utilize an independent [CLS] token to extract global features by interacting with other image patch tokens [130, 123, 127]. Only the [CLS] token is used for classification, which prevents the image tokens from being encoded too abstractly. Additionally, feeding the classification results into the text decoder improves the quality of the generated reports. [147, 44, 86, 130, 123, 54, 133].

A single [CLS] token may not accurately cover all diseases, similar to how a general practitioner's diagnosis may not be as precise as that of a specialist. METransformer [128] addresses this by concatenating multiple expert tokens in the image encoder and using orthogonal loss to minimize overlap among these tokens, thereby encouraging them to capture complementary information. The model generates a report based on each expert token and selects the best report through a voting strategy.

3.2.2 Detection and Segmentation

In addition to using classification tasks to guide the image encoder towards clinically relevant areas, researchers have proposed using pre-trained segmentation or detection networks to explicitly assist the encoder in targeting anatomical regions [158, 156, 126, 116]. One approach employs SAM [61] to segment meaningful anatomical regions (e.g., the left and right lungs) from chest X-ray images before inputting them into the image encoder [158]. This method eliminates background interference, thereby enhancing the encoder's focus on relevant regions. Other researchers combine global features extracted by the image encoder with regional features from the detection or segmentation network, inputting both into the decoder to enrich the information it receives [156, 126]. In addition, because lesions typically occupy a small portion of medical images, Tanida *et al.* [116] propose a framework that compels the model to focus on the critical regions. This framework involves cropping multiple anatomical regions from the input image using a detection network, followed by multiple binary classification networks to evaluate whether each region is critical for report generation. The text decoder processes only the critical regions, thereby preventing it from being overwhelmed by the numerous normal regions.

3.2.3 Internal Structure of Image Encoder

In addition to adopting a joint learning strategy and using pre-trained auxiliary networks, modifying the internal structure of the image encoder can also enhance its focus on lesion areas. Two effective and widely used methods are cross-attention and high-order attention.

The cross-attention mechanism [118] assigns weights to image regions based on their relevance to disease tags, thereby emphasizing the features of regions containing lesions. This method does not require disease annotations for each image, but rather a set of all disease tags [16, 147, 83]. Specifically, the disease tag set is used as the query, and the image is used as the key and value. The dot product in the cross-attention mechanism can select disease tags related to the image content and enhance the features of regions containing these diseases.

Recently, several studies have attempted to replace traditional first-order attention with X-linear attention [98] in Transformer-based image encoders [130, 128, 137, 124]. X-linear attention captures complex high-order interactions within medical images, leading to a more nuanced and comprehensive understanding of the images and more accurate localization of abnormalities. In detail, given a query $Q \in \mathbb{R}^{D_q}$, a set of keys $K = \{k_i\}_{i=1}^{N}$ and a set of values $V = \{v_i\}_{i=1}^{N}$, where $k_i \in \mathbb{R}^{D_k}$ and $v_i \in \mathbb{R}^{D_v}$, low-rank bilinear pooling [59] is performed to obtain the joint bilinear query-key B_k and query-value B_v:

$$B_i^k = \sigma(W_k k_i) \odot \sigma(W_q^k Q), \quad B_i^v = \sigma(W_v v_i) \odot \sigma(W_q^v Q), \tag{12}$$

where $W_k \in \mathbb{R}^{D_B \times D_k}$, $W_v \in \mathbb{R}^{D_B \times D_v}$, and $W_q^k, W_q^v \in \mathbb{R}^{D_B \times D_q}$ are learnable matrices, σ denotes ReLU unit, and $\odot$ represents element-wise multiplication. Then, the spatial attention β_i^s and channel-wise attention β^c are computed as follows:

$$\begin{aligned} B_i^{'k} &= \sigma(W_B^k B_i^k), \quad \beta_i^s = softmax(W_s B_i^{'k}) \\ \bar{B} &= \frac{1}{N}\sum_{i=1}^{N} B_i^{'k}, \quad \beta^c = sigmoid(W_c \bar{B}), \end{aligned} \tag{13}$$

where $W_B^k \in \mathbb{R}^{D_c \times D_B}$, $W_s \in \mathbb{R}^{1 \times D_c}$, and $W_c \in \mathbb{R}^{D_B \times D_c}$ are learnable matrices. Finally, the output of the X-linear attention mechanism is given by:

$$\hat{v} = F_{X-linear}(K, V, Q) = \beta^c \odot \sum_{i=1}^{N} \beta_i^s B_i^v \tag{14}$$

3.3 Enhancing Text Decoder With Supplementary Information

This section presents three approaches for augmenting the text decoder with supplementary information: (i) retrieving similarity reports (Section 3.3.1), (ii) leveraging memory (Section 3.3.2), and (iii) integrating knowledge graphs (Section 3.3.3). Each approach addresses specific challenges, such as ensuring clinical consistency, alleviating privacy concerns, and building medical knowledge for better model comprehension. Figure 6 shows the flowcharts of the three methods.

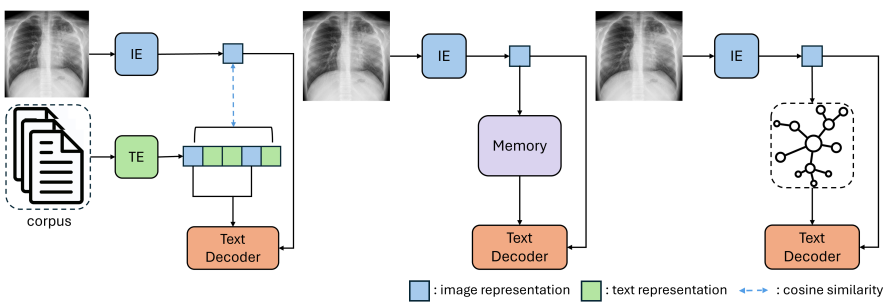

Figure 6: Flowcharts of three representative methods for augmenting the text decoder with supplementary information. The left diagram shows a retrieval-based approach. Reports similar to the input image are found from the corpus (consisting of training reports) based on cosine similarity and are input into the text decoder as reference information. The middle diagram illustrates replacing the corpus with a learnable memory to avoid leakage of training data. The right diagram demonstrates replacing the memory with a knowledge graph to store the clinical information used to generate the report in a structured manner. IE and TE represent image encoder and text encoder, respectively.

3.3.1 Retrieve Similarity Reports

Given the limited diversity of diagnoses in medical reports, a large retrieval corpus can adequately cover the potential diagnoses of input images. Some researchers have proposed using a retrieval-based approach to generate new reports, with the primary advantage being the clinical consistency of the generated reports with manually written ones [30, 52]. To elaborate, image and text encoders are trained using the CLIP method, which produces higher similarity scores for paired image-text examples and lower scores for unpaired ones. A corpus is then constructed using the reports in the training set. During inference, the model retrieves the top K reports from the corpus with the highest similarity scores to the input image and combines them into the predicted report. However, since candidate selection is based on maximizing similarity scores, the predicted report is prone to repeating information.

PPKED [83] improves the basic retrieval-based approach by modifying the retrieval process and using a text decoder to generate reports instead of merely combining retrieved candidates. The corpus is constructed using image-text pairs from the training set. During inference, the system retrieves the top K images in the corpus that are most similar to the input image and uses their corresponding reports to enhance the image features. Finally, the text decoder generates the final report based on the enhanced image features, ensuring coherence and the absence of redundant content.

3.3.2 Memory

However, retrieving training data during inference raises concerns regarding the privacy of medical data. Some researchers have proposed a solution by employing learnable memory to replace the corpus [141, 24]. The memory stores features derived from the training data, rather than the training data itself, thereby mitigating the risk of data leakage. Yang *et al.* [141] use cross-attention to update the memory during training and to enhance image features during inference. Cheng *et al.* [24] adopt a more explicit approach to update the sentence-prototype memory. During training, the sentence prototype most similar to the input sentence is selected using cosine similarity, and the memory is updated based on the L1 loss between the prototype and the input sentence.

Furthermore, integrating memory into the text decoder is another method to enhance the quality of the generated reports [85, 130, 23, 125]. This memory records fine-grained medical knowledge and historical information from previous generation processes, which is valuable for generating lengthy texts. One approach involves using the memory matrices to augment the keys and values of the Transformer-based decoder [85, 130]. Specifically, given a key K and value V, the memory-augment key and value are defined as $\hat{K} = [K, M_k]$ and $\hat{V} = [V, M_v]$, where M_k and M_v are learnable matrices, and

$[\cdot, \cdot]$ denotes concatenation. Another approach utilizes a gate mechanism to update the memory and map it to the scale and offset parameters in layer normalization, thereby injecting the memory into the decoder [23, 125].

3.3.3 Knowledge Graph

A more structured memory, in the form of a medical knowledge graph, can group diseases according to organs or body parts. This is because abnormalities in the same body part often exhibit strong correlations and common features. Initially, a medical graph is designed based on prior knowledge from chest findings to cover common abnormalities and their relationships [153]. In this graph, disease keywords serve as nodes (V) and their relationships as edges (E), denoted as $G = \{V, E\}$. Graph convolution [60] is used to propagate information within the graph, thereby enhancing the model's capacity to comprehend medical knowledge. This pre-constructed graph has been adopted by several studies, which further distill the knowledge graph during the decoding stage to enrich the information received by the decoder [152, 46, 83].

However, a fixed graph may not contain all the necessary knowledge about the input image, thereby limiting its effectiveness. Li *et al.* [72] design a dynamic graph that uses the nodes from the pre-constructed graph [153] as initial nodes and models relationships with an adjacency matrix. In this matrix, 1 represents a connection between two nodes, while 0 indicates no relationship. The dynamic update process is as follows: during training, the model retrieves the top three most similar reports from a corpus for each input image. Then, RadGraph [51] is applied to extract specific knowledge triplets from these reports, formatted as {*subject entity*, *relation*, *object entity*}. If only the subject or object entity is present in the graph, the other entity in the triplet is added as an additional node, and their relation is set to 1 in the adjacency matrix, indicating a link between the two nodes.

In addition to learning the relationships between entities, MGSK [142] uses more explicit and accurate relationships that are manually annotated. The model comprises two graphs: a general graph and a specific graph. The general graph is independent of the input image and is manually constructed by radiologists from 500 radiology reports in the MIMIC-CXR dataset [51]. The general knowledge is stored in the triplet, {*subject entity*, *relation*, *object entity*}, where relations include 'suggestive of,' 'modify,' and 'located at'. The specific knowledge is retrieved from the corpus by finding the top ten images most similar to the input image and extracting triplets from their corresponding reports using RadGraph. The general and specific knowledge graphs are fused with image features to enrich the input of the text decoder.

3.4 Refining Generated Reports

This section outlines methods designed to refine the accuracy and semantic coherence of generated medical reports. In particular, these techniques are designed to guarantee that the generated content accurately reflects essential medical insights that are critical for clinical reliability. Two pivotal approaches will be discussed: (i) the traceback mechanism (Section 3.4.1) , which evaluates semantic fidelity, and (ii) reinforcement learning (Section 3.4.2), which correlates training goals with evaluative metrics.

3.4.1 Traceback Mechanism

Most medical report generation methods construct loss functions that evaluate the discrepancy between generated and GT reports at the word level (for further details, please refer to Equation 2). Consequently, models tend to predict frequently observed words in order to achieve a high overlap rate [41], which may result in the generation of clinically flawed reports. A high-quality generated report should also be semantically similar to the GT. To achieve this, some researchers propose a traceback mechanism to control the semantic validity of generated content through self-assessment [131, 65, 145, 20]. This approach involves inputting the generated report T into a text encoder to extract semantic features x_t, and optimizing the model to ensure that x_t is similar to the semantic features x_{t^*} of the GT report T^*. The process of the traceback mechanism is shown in Figure 7.

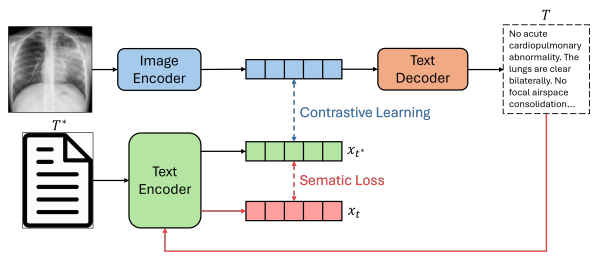

Figure 7: The GT report T^* is input into the text encoder to obtain semantic features x_{t^*} (text representation). The visual representation extracted by the image encoder is then fed into the text decoder after contrastive learning, producing the generated report T. The traceback mechanism begins by inputting this generated report T back into the text encoder to extract the semantic features x_t. The difference between x_{t^*} and x_t, termed the semantic loss, serves as the objective function of the traceback mechanism. This mechanism aims to reduce the discrepancy between the generated report T and the GT report T^* at the feature level.

A variety of techniques exist for measuring semantic loss, including calculating the L2-norm distance [131] and the cosine similarity [65] between x_t

and x_{t^*}, as well as using a classifier to ensure that the disease classifications based on x_t and x_{t^*} are the consistent [145]. Classifying diseases based on the semantic features is a preferable approach because it encourages the model to correct generated words that influence disease classification, which is the most critical aspect of medical reports. A more complex approach involves having the model synthesize a medical image I based on the generated report T and comparing I with the input image I^* [20]. Moreover, to reduce significant gradient fluctuations during the initial training stages, it is advantageous to assign a smaller weight to the traceback semantic loss at the beginning [145].

3.4.2 Reinforcement Learning

In contrast to indirect semantic loss, some researchers use reinforcement learning with NLP metrics as rewards to align training goals with final evaluation criteria [77, 131, 130, 25, 102, 80]. Specifically, the text decoder is treated as "agent" that interacts with an external "environment" (visual and textual features). The network parameters, θ, define a "policy" p_θ, that results in an "action" (the prediction of the next word). The CIDEr score is used as a reward r, which is calculated by comparing the generated sequence to the corresponding GT sequence. The objective of training is to minimize the negative expected reward:

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)], \tag{15}$$

where $w^s = (w_1^s, ..., w_T^s)$ and w_t^s represents the word sampled from the model at the time step t. The expected gradient of the non-differentiable reward function can be approximated using a Monte-Carlo sample $w^s = (w_1^s, ..., w_T^s)$ from p_θ:

$$\nabla_\theta L(\theta) \approx -(r(w^s) - r(\hat{w}))\nabla_\theta \log p_\theta(w^s), \tag{16}$$

where $r(\hat{w})$ is the reward obtained by the current model under the inference algorithm at test time, and $\hat{w}$ is generated by greedy decoding:

$$\hat{w}_t = \arg\max_{w_t} p(w_t|\hat{w}_0, ..., \hat{w}_{t-1}, I), \tag{17}$$

where $\hat{w}_0, ..., \hat{w}_{t-1}$ are the previous generated words and I is the image representation. As a result, samples w^s from the model that yield a higher reward than $\hat{w}$ increase their probability during the learning process, whereas samples resulting in a lower reward are suppressed. The process of implementing reinforcement learning is shown in Figure 8.

However, using only one NLP metric (e.g. CIDEr) as a reward may lead to partial optimization rather than overall optimization, as long text generation tasks cannot rely on a single metric to evaluate performance. Xu et al. [137] test seven NLP metrics with different combinations as rewards and find that

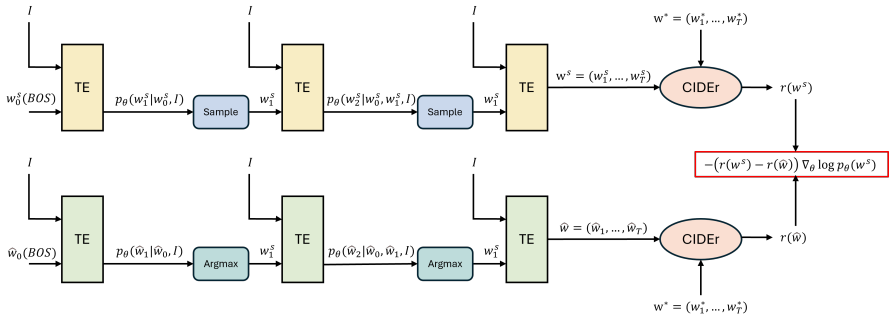

Figure 8: Implementation of reinforcement learning involves several key elements. I represents the image representation, while BOS is a special token denoting the beginning of the sequence. w^s is the sequence obtained by sampling, $\hat{w}$ is the sequence obtained by the argmax operation, and w^* is the GT sequence. By calculating the CIDEr scores for w^s and w^*, as well as for $\hat{w}$ and w^*, the rewards $r(w^s)$ and $r(\hat{w})$ are obtained. The gradient of the objective function, highlighted in the red box, is then computed based on these rewards. TE represents text decoder.

using BLEU-4, METEOR, and CIDEr together achieves the best results. Miura *et al.* [89] design the factual completeness and consistency rewards that are more suitable for medical reports. These special rewards serve to ascertain whether the generated report and the GT report contain the same anatomical entities and whether the sentences containing these entities contradict the corresponding sentences in the GT report.

3.5 Dataset Limitations

Medical image-text paired datasets are constrained by two fundamental limitations: noise and limited size. Noise arises from temporal information and false negatives, which can distort the training data and lead to inaccurate model predictions. Temporal information noise occurs when reports reference earlier images, which introduces inconsistencies that models struggle to interpret correctly. False negatives, on the other hand, arise in contrastive learning when similar reports are incorrectly treated as negative samples, confusing the model. Additionally, the limited size of these medical datasets presents another challenge, as insufficient data hinders the model's capacity to generalize and perform effectively. The following subsubsections present innovative methods to address these issues: (i) removing temporal noise and mitigating false negatives in contrastive learning (Section 3.5.1), (ii) using LLM to improve model performance on limited datasets (Section 3.5.2), and (iii) expanding training datasets through semi-supervised and unsupervised learning (Section 3.5.3).

3.5.1 Eliminating Noise

Although medical image-report datasets authorized by professionals are generally more accurate than general image-text datasets in terms of annotation, they still exhibit inherent noise, including temporal information and false negatives. Temporal information noise can be attributed to reports that reference earlier images. For instance, in a report, "**Comparison made to prior study**, there is **again** seen moderate congestive heart failure with **increased** vascular cephalization, **stable**. There are large bilateral pleural effusions **but decreased since previous**", the bolded words relate to earlier images, yet the current image paired with the report does not contain this comparative information. Such references to previous data introduce temporal noise, which may lead trained models to generate hallucinations about non-existent priors. False negatives occur in contrastive learning when negative or unpaired texts (reports from other patients) describe identical symptoms as the paired reports. Simply treating the other reports as negative samples introduces noise into the supervision process, thereby confusing the model.

To eliminate temporal noise, Ramesh *et al.* [104] propose two approaches: rewriting medical reports using GPT-3 and designing a token classifier to delete word tokens associated with prior studies. The latter method is more accurate and cost-effective. In contrast to eliminating temporal information, some methods effectively integrate it into the learning process [9, 43]. For example, the BioViL-T framework [9] assumes that a patient has a current image I_c, a current report T_c, and a previous image I_p. The images I_c and I_p are processed through a CNN to generate features P_c and P_p, respectively. These features are then fed into Transformer blocks to extract difference features P_d. Subsequently, P_c and P_d are input together into the text decoder. If the model's input includes only the current image I_c, P_d is substituted with a learnable feature P_m.

To address false negatives within contrastive learning, ALBEF [69] utilizes momentum distillation to mitigate the impact of such noise in general image captioning datasets. The momentum encoder, functioning as a teacher, produces a stable set of pseudo-labels for the input image. These pseudo-labels serve as training targets for the student encoder to account for the potential positives in the negative pairs. Research has shown that this momentum distillation can be seamlessly adapted to medical datasets [72, 52]. Furthermore, MedCLIP [132] decouples medical image-text pairs and employs semantic similarity to create pseudo-labels. This approach entails extracting entities from the reports as textual labels and utilizing disease labels as image labels. The cosine similarity between these image and text labels serves as pseudo-labels in contrastive learning.

3.5.2 Large Language Model

To address the issue of the limited size of medical image-report datasets, a feasible approach is to utilize a pre-trained large language model (LLM) as a text decoder. This method leverages the LLM's robust language generation and zero-shot transfer capabilities, thereby reducing the number of parameters that need to be trained from scratch. Li *et al.* [67] propose a trainable mapping module, Q-Former, to bridge the gap between a frozen image encoder and a frozen LLM. Specifically, they employ a pre-trained Vision Transformer (ViT) as an image encoder to extract image features, which are then mapped to the text feature space by Q-Former. Subsequently, the frozen LLM serves as a text decoder to generate reports. MSMedCap [156] demonstrates that Q-Former is effective for handling limited medical image-report data.

Nevertheless, utilising a frozen LLM as a text decoder concurrently with a frozen image encoder is not the optimal approach. Research indicates that fine-tuning both the image encoder and the mapping module when the LLM decoder is frozen can enhance the quality of the generated reports [129]. Additionally, freezing the LLM may not be the most effective strategy, as an LLM pre-trained on general data may not be suitable for the medical domain. It is therefore generally recommended to fine-tune the LLM on task-specific data. However, it should be noted that fine-tuning an LLM requires a substantial amount of data, so directly fine-tuning an LLM with limited medical data may result in suboptimal performance.

Liu *et al.* [81] propose a coarse-to-fine decoding strategy to fine-tune an LLM on limited medical datasets in a bootstrapping manner. They initially employ MiniGPT-4 [159] to generate a coarse report and then use the coarse report as a prompt, along with the image features, to input into the decoder of MiniGPT-4 again to generate the final refined report. Additionally, pseudo self-attention [160] is another method for fine-tuning an LLM on medical data [116, 4]. This approach introduces new parameters solely within the self-attention block, while other parameters of the Transformer are initialized with pre-trained values. Given an image feature X and a hidden state Y, the pseudo self-attention is formulated as follows:

$$PSA(X, Y) = softmax((YW_q)\begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^T)\begin{bmatrix} XU_v \\ YW_v \end{bmatrix}, \tag{18}$$

where U_k, U_v are new parameters, and W_q, W_k, W_v are parameters from pre-trained model. Fine-tuning an LLM with pseudo self-attention results in minimal changes to the pre-trained LLM's parameters, thereby maintaining its text generation capabilities [160].

3.5.3 Semi-Supervised and Unsupervised Learning

To address the limited size of medical image-text paired datasets, some researchers have explored semi-supervised and unsupervised learning methods to expand the training dataset. The RAMT model [152] employs a student-teacher network to train in a semi-supervised manner using 25% paired data and 75% unpaired image data. The student and teacher networks share the same structure but have distinct parameters. During the training phase, the teacher network parameters are updated by the exponential moving average (EMA) of the student network parameters. Different noises are added to the input images, which are then fed into the student and teacher networks, respectively. The output of the teacher network serves as supervision for the student network.

However, the semi-supervised method still requires some images with corresponding reports. KGAE [85] addresses this limitation by utilizing unsupervised learning. The model employs a pre-constructed knowledge graph G as a shared latent space to bridge the gap between image and text representations. Given an input image I and an input report R, the graph G maps them into the same latent space, G_I and G_R. For unsupervised learning, the image encoder and text decoder are trained separately. To train the image encoder, G_I and G_R jointly implement disease classification, ensuring they form a common latent space. To train the text decoder, the report R is reconstructed from G_R, following the process $R \rightarrow G_R \rightarrow R$. Additionally, KGAE can be applied in semi-supervised and supervised settings. In these settings, as well as for inference, the pipeline follows $I \rightarrow G_I \rightarrow R$.

In a more recent study, Hirsch *et al.* [40] refine the unsupervised method, achieving higher accuracy than KGAE. The method employs cycle consistency to ensure that cross-modal mappings retain information. Cross-modal mappings include image-to-report ($I2R$) and report-to-image ($R2I$). The objective of cycle consistency is to minimise the differences between z_i and $R2I(I2R(z_i))$, as well as z_r and $I2R(R2I(z_r))$, where z_i and z_r represent image and text representations, respectively. To ensure that the image encoder extracts relevant semantic information (e.g., diseases and organs), they employ contrastive learning to align the image representations (output by the image encoder) with the text representations of pseudo-reports. These pseudo-reports are constructed based on the disease labels of images. To train the text decoder, they also adopt a report reconstruction task and add adversarial learning to ensure that the text decoder receives similar features during training and inference.

4 Applications

In this section, we introduce the applications of AMRG across various medical imaging modalities, including chest radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, ophthalmic imaging, endoscopy, surgical scene, and pathological imaging. These modalities are crucial for diagnosing a broad spectrum of medical conditions.

Chest Radiography: In recent years, a significant amount of research focused on chest radiography report generation (refer to Table 1). This is largely due to the availability of large and publicly accessible datasets such as MIMIC-CXR [56] and IU X-ray [27], which contain extensive collections of annotated images and paired reports. The availability of these datasets enables the models to effectively learn the intricate relationships between visual features and textual descriptions.

Table 1: Comparison of radiographic report generation models (2021–2024) on benchmark datasets. The 'Code' column shows if the code is publicly accessible, and the 'Method' column lists the techniques used. Techniques mentioned in Section 3 follow the names in Figure 2, while others retain their original names (e.g., curriculum learning). The highest and second-highest values in each column are bolded and underlined, respectively. Abbreviations: BL-4 (BLEU-4), MTR (METEOR), RG-L (ROUGE-L), CD (CIDEr), P (Precision), R (Recall), and F (F1 Score).

Model	Year	Code	MIMIC-CXR							IU-Xray				Method	Dataset
			BL-4	MTR	RG-L	CD	P	R	F	BL-4	MTR	RG-L	CD		
[83]	2021		0.106	0.149	0.284	–	–	–	–	0.168	–	0.376	0.351	Encoder structure, retrieve similarity reports, knowledge graph	MIMIC-CXR, IU-Xray
[139]	2021		0.107	0.144	0.274	–	0.385	0.274	0.294	–	–	–	–	Global alignment	MIMIC-CXR, MIMIC-ABM
[84]	2021		0.109	0.151	0.283	–	0.352	0.298	0.303	0.169	0.193	0.381	–	Contrastive attention	MIMIC-CXR, IU-Xray
[147]	2021		0.112	0.158	0.283	–	–	–	–	0.173	0.204	0.379	–	Disease classification, encoder structure	MIMIC-CXR, IU-Xray
[89]	2021	✓	0.114	–	–	**0.509**	0.503	0.651	0.567	0.131	–	–	**1.034**	Reinforcement learning	MIMIC-CXR, IU-Xray
[85]	2021		0.118	0.153	0.295	–	0.389	0.362	0.355	0.179	0.195	0.383	–	Disease classification, memory, knowledge graph, unsupervised learning	MIMIC-CXR, IU-Xray
[144]	2021	✓	0.143	–	0.326	0.273	0.237	0.326	–	0.180	–	0.398	0.439	Global alignment, disease classification	MIMIC-CXR, IU-Xray
[93]	2021	✓	**0.224**	**0.222**	**0.390**	–	0.432	0.418	0.412	**0.235**	0.219	**0.436**	–	Human-computer interaction, disease classification, memory, traceback mechanism	MIMIC-CXR, IU-Xray
[131]	2021		–	–	–	–	–	–	–	0.208	–	0.359	0.452	Global alignment, traceback mechanism, reinforcement learning	COV-CTR, IU-Xray
[42]	2021	✓	–	0.101	0.240	0.493	–	–	–	–	–	–	–	Encoder structure	MIMIC-CXR, IU-Xray
[4]	2021	✓	–	–	–	–	–	–	–	0.111	0.164	0.289	0.257	Large language model	MIMIC-CXR, IU-Xray
[82]	2022		0.097	0.133	0.281	–	–	–	–	0.162	0.186	0.378	–	Curriculum learning	MIMIC-CXR, IU-Xray
[121]	2022	✓	0.105	0.138	0.279	–	–	–	–	0.199	0.22	0.411	0.359	Intermediate matrix	MIMIC-CXR, IU-Xray
[22]	2022	✓	0.106	0.142	0.278	–	0.334	0.275	0.278	0.170	0.191	0.375	–	Intermediate matrix	MIMIC-CXR, IU-Xray

Table 1: Continued.

Model	Year	Code	MIMIC-CXR							IU-Xray				Method	Dataset
			BL-4	MTR	RG-L	CD	P	R	F	BL-4	MTR	RG-L	CD		
[102]	2022	✓	0.109	0.151	0.287	–	0.342	0.294	0.292	0.181	0.201	0.384	–	Intermediate matrix, reinforcement learning	MIMIC-CXR, IU-Xray
[142]	2022	✓	0.115	–	0.284	0.203	–	–	–	0.178	–	0.381	0.382	Knowledge graph	MIMIC-CXR, IU-Xray
[127]	2022		0.118	–	0.287	0.281	–	–	–	0.175	–	0.377	0.449	Global alignment, disease classification	MIMIC-CXR, IU-Xray
[123]	2022	✓	0.121	0.147	0.284	–	–	–	–	0.188	0.208	0.382	–	Disease classification	MIMIC-CXR, IU-Xray
[130]	2022	✓	0.136	0.170	0.298	0.429	–	–	–	–	–	–	–	Disease classification, encoder structure, memory, reinforcement learning	MIMIC-CXR
[92]	2022	✓	0.136	0.191	0.315	–	0.396	0.312	0.350	0.170	0.230	0.390	–	Global alignment	MIMIC-CXR, IU-Xray
[148]	2022		–	–	–	–	–	–	–	0.174	0.193	0.377	–	Intermediate matrix, disease classification	IU-Xray
[120]	2022		–	–	–	–	–	–	–	0.175	–	0.36	0.331	Disease classification	CX-CHR, IU-Xray
[65]	2022		–	–	–	–	–	–	–	0.215	0.201	0.415	–	Traceback mechanism	IU-Xray
[73]	2023	✓	–	–	–	–	–	–	–	0.125	–	0.279	0.306	Disease classification, detection, knowledge graph	CX-CHR, COV-CTR, IU-Xray
[87]	2023		0.069	–	0.235	–	–	–	0.32	–	–	–	–	Large language model	MIMIC-CXR
[9]	2023	✓	0.092	–	0.296	–	–	–	–	–	–	–	–	Global alignment, local alignment, eliminate noise	MIMIC-CXR, MS-CXR-T
[134]	2023		0.103	0.139	0.270	0.109	–	–	–	0.18	0.206	0.369	0.287	Global alignment	MIMIC-CXR, IU-Xray
[75]	2023		0.107	0.157	0.289	0.246	–	–	–	0.200	0.218	0.405	0.501	Global alignment, intermediate matrix	MIMIC-CXR, IU-Xray
[72]	2023	✓	0.109	0.150	0.284	0.281	–	–	–	0.163	0.193	0.383	0.586	Global alignment, knowledge graph, eliminate noise	MIMIC-CXR, IU-Xray
[141]	2023	✓	0.111	–	0.274	0.111	0.420	0.339	0.352	0.174	–	0.399	0.407	Disease classification, memory	MIMIC-CXR, IU-Xray
[151]	2023		0.113	0.143	0.276	–	–	–	–	0.190	0.207	0.394	–	Memory	MIMIC-CXR, IU-Xray, COV-CTR
[46]	2023		0.113	0.160	0.285	–	0.371	0.318	0.321	0.185	0.242	0.409	–	Disease classification, knowledge graph	MIMIC-CXR, IU-Xray
[16]	2023		0.116	0.161	0.283	–	–	–	–	0.175	–	0.375	0.361	Encoder structure, memory	MIMIC-CXR, IU-Xray
[126]	2023		0.118	0.136	0.301	–	–	–	–	0.176	0.205	0.396	–	Detection	MIMIC-CXR, IU-Xray
[125]	2023		0.119	0.158	0.286	0.259	–	–	–	0.205	0.223	0.414	0.370	Local alignment	MIMIC-CXR, IU-Xray
[44]	2023	✓	0.123	0.162	0.293	–	0.416	0.418	0.385	0.195	0.205	0.399	–	Disease classification	MIMIC-CXR, IU-Xray
[128]	2023		0.124	0.152	0.291	0.362	–	–	–	0.172	0.192	0.380	0.435	Encoder structure	MIMIC-CXR, IU-Xray
[43]	2023	✓	0.125	0.168	0.288	–	0.389	0.443	0.393	–	–	–	–	Disease classification, eliminate noise	MIMIC-CXR, MIMIC-ABN
[116]	2023	✓	0.126	0.168	0.264	0.495	–	–	–	–	–	–	–	Detection, disease classification, large language model	MIMIC-CXR
[95]	2023	✓	0.127	0.155	0.286	0.389	0.367	0.418	0.391	0.175	0.200	0.376	0.694	Warm starting	MIMIC-CXR, IU-Xray
[112]	2023		0.130	0.148	0.315	–	–	–	–	0.174	–	0.388	–	Encoder structure, memory	MIMIC-CXR, IU-Xray
[137]	2023		0.192	0.207	0.380	0.372	–	–	–	0.149	0.197	0.381	0.524	Encoder structure, reinforcement learning	MIMIC-CXR, IU-Xray
[129]	2023	✓	0.134	0.160	0.297	0.269	0.392	0.387	0.389	0.173	0.211	0.377	0.438	Large language model	MIMIC-CXR, IU-Xray
[152]	2023		0.113	0.153	0.284	–	0.380	0.342	0.335	0.165	0.195	0.377	–	Knowledge graph, semi-supervised learning	MIMIC-CXR, IU-Xray

Table 1: Continued.

Model	Year	Code	MIMIC-CXR							IU-Xray				Method	Dataset
			BL-4	MTR	RG-L	CD	P	R	F	BL-4	MTR	RG-L	CD		
[86]	2023		0.125	0.160	0.304	–	**0.855**	**0.730**	**0.773**	0.206	0.211	0.423	–	Human-computer interaction, global alignment, disease classification	MIMIC-CXR, IU-Xray
[140]	2023		–	–	0.225	0.160	–	–	–	–	–	0.341	0.380	Knowledge graph, disease classification	MIMIC-CXR, IU-Xray
[158]	2023		–	–	–	–	–	–	–	0.221	0.210	0.433	–	Global alignment, segmentation	IU-Xray
[124]	2023		–	–	–	–	–	–	–	0.157	0.196	0.374	–	Encoder's structure	IU-Xray
[20]	2024		0.106	0.163	0.286	–	–	–	–	0.145	0.162	0.366	–	Encoder's structure, reinforcement learning	MIMIC-CXR, IU-Xray
[54]	2024	✓	0.112	0.157	0.268	–	0.501	0.509	0.476	0.098	0.160	0.281	–	Retrieve similarity reports, disease classification	MIMIC-CXR, IU-Xray
[122]	2024	✓	0.112	0.145	0.279	0.161	0.483	0.323	0.387	0.218	0.203	0.404	0.418	Disease classification	MIMIC-CXR, IU-Xray
[117]	2024		0.115	–	0.275	–	–	–	0.398	–	–	–	–	Disease classification	MIMIC-CXR
[34]	2024	✓	0.116	0.168	0.286	–	0.482	0.563	0.519	0.205	0.210	0.409	–	Retrieve similarity reports, memony	MIMIC-CXR, IU-Xray
[146]	2024	✓	0.121	0.149	0.281	–	0.319	0.509	0.393	0.194	0.218	0.402	–	Reinforcement learning	MIMIC-CXR, IU-Xray
[97]	2024		0.125	0.154	0.291	–	–	–	–	0.172	0.206	0.401	–	Intermediate matrix	MIMIC-CXR, IU-Xray
[81]	2024	✓	0.128	0.175	0.291	–	0.465	0.482	0.473	0.184	0.208	0.390	–	Retrieve similarity reports, large language model	MIMIC-CXR, IU-Xray
[145]	2024		0.129	0.162	0.309	0.311	–	–	–	0.204	**0.233**	0.386	0.469	Local alignment, disease classification, traceback mechanism	MIMIC-CXR, IU-Xray
[80]	2024		0.141	0.163	0.309	–	0.457	0.337	0.330	0.175	0.192	0.379	0.368	Local alignment, reinforcement learning	MIMIC-CXR, IU-Xray
[40]	2024		0.072	0.128	0.239	–	0.237	0.197	0.183	0.140	0.197	0.360	–	Unsupervised learning	MIMIC-CXR, IU-Xray, PadChest

The application of AMRG in radiography offers several significant benefits. First, it can significantly reduce the radiologist's workload by automating the initial draft of the report, allowing them to focus on more complex and nuanced cases [4, 134, 3, 29, 113]. Second, these models can improve diagnostic consistency and reduce inter-observer variability by applying standardized criteria and guidelines in the report generation process [150]. Third, in regions with limited access to experienced radiologists, AMRG models can provide essential diagnostic support, ensuring timely and accurate medical care for patients [91]. Finally, these models can support large-scale screening programs by rapidly processing and generating reports for large volumes of X-ray images, thereby facilitating the early detection of diseases.

3D Imaging: CT and MRI provide detailed, three-dimensional (3D) views of the human body, playing a pivotal role in diagnosing a wide range of conditions, including neurological disorders and abdominal diseases. Recent studies have explored AMRG for these imaging modalities [38, 21, 151], but these studies often treat 3D images as a set of 2D slices, overlooking the inherent stereoscopic structural information, an issue that future research should address.

Ultrasound: Ultrasound is widely used due to its real-time imaging capability and safety profile. Recently, AMRG has been explored for ultrasound applications [66]. enabling real-time report generation and assisting clinicians in making immediate decisions, especially in emergency and point-of-care settings. However, the low image quality and operator-dependent nature of ultrasound image acquisition still affect the quality of generated reports.

Ophthalmic Imaging: In ophthalmology, AMRG applications in fundus fluorescein angiography (FFA) [71] and fundus images [133] aid in diagnosing critical eye diseases such as diabetic retinopathy. Li *et al.* [71], proposed the CGT model for ophthalmic report generation. Their approach involves an information extraction scheme that converts unstructured medical reports into a structured format, constructing clinical graphs. These graphs encapsulate prior medical knowledge, which is then distilled into sub-graphs and integrated with visual features to enhance report generation. By employing a combination of cross-entropy and triples loss, they optimize the report generation model, achieving SOTA results on the FFA-IR benchmark dataset [70].

Endoscopy: For endoscopic imaging, AMRG aids in diagnosing various complications, such as gastrointestinal diseases and cancers. Cao *et al.* [16] combined disease tags with cross-attention and introduced memory augmentation in the image encoder to improve the model's sensitivity to lesion areas. Their model achieved competitive results with SOTA models on the gastrointestinal endoscope image dataset, which is a private dataset contains white light images and their Chinese reports from the department of gastroenterology.

Surgical Scene: In surgical imaging, AMRG helps create operative reports by documenting surgical steps. This alleviates surgeons' workloads and allows them to focus more on the operations. Lin *et al.* [77] proposed the SGT++ model to effectively models interactions between surgical instruments and tissues. Their method involves homogenizing heterogeneous scene graphs to learn explicit, structured, and detailed semantic relationships via an attention-induced graph Transformer. Additionally, it incorporates implicit relational attention to integrate prior knowledge of interactions.

Pathological Imaging: Pathological imaging involves examining high-resolution images of tissue sections, requiring meticulous analysis. Chen *et al.* [19] recently introduced the MI-Gen model to produce pathology reports for gigapixel whole slide images (WSIs). Furthermore, they created the largest WSI-text dataset, PathText, which contains nearly 10,000 high-quality WSI-text pairs.

Unified Model: Unlike the above models specialized for one modality, Google recently introduced a groundbreaking model, Med-PaLM M [117], which encodes and interprets various biomedical data modalities using the same model weights. This model can process multiple data modalities, including clinical language, genomics, and imaging (e.g., radiography, mammography, dermatology, and pathology). To support these developments, they curated

MultiMedBench [117], a benchmark comprising 14 tasks such as AMRG, report summarization, medical question answering, visual question answering, and medical image classification. Med-PaLM M achieved performance competitive with or surpassing specialist models on all MultiMedBench tasks. This innovation marks a significant advance in applying a unified model to different modalities.

5 Public Dataset

In this section, we introduce several image-report datasets used for AMRG. All discussed datasets are publicly available and privacy-safe. The two benchmark datasets (Section 5.1) are widely used and serve as standards for performance comparison in most AMRG studies. Additionally, other datasets (Section 5.2) have been created to address specific needs, such as particular languages and imaging modalities. Table 2 presents the statistical results of the datasets. The following sections introduce these datasets in detail.

Table 2: This table presents dataset statistics including counts of images, reports, abnormal/normal cases, and medical conditions. For MIMIC-CXR, abnormal/normal counts refer to the counts of radiographs classified by Ni *et al.* [94], as official splits aren't provided. For other datasets, abnormal/normal counts are provided by officials and represent the counts of reports (cases).

Dataset	Images	Reports	Abnormal	Normal	Conditions
Chest radiographs					
IU-Xray [27]	7,470	3,955	2,470 (62.5%)	1,485 (37.5%)	177
MIMIC-CXR [56]	377,110	227,835	38,551 (10.2%)	338,559 (89.8%)	14
Padchest [15]	160,868	109,931	-	-	19
CX-CHR [76]	45,598	33,236	-	-	20
Lung CT scans					
COV-CTR [73]	728	728	349 (47.9%)	379 (52.1%)	2
Fundus fluorescein angiography images					
FFA-IR [70]	1,048,584	10,689	10,087 (94.4%)	602 (5.6%)	46
Surgical images					
EndoVis-18 [135, 136]	1,560	1,560	-	-	20
TORS [135, 136]	335	335	-	-	13

5.1 Benchmark Datasets

IU-Xray: The Indiana University Chest X-ray dataset (IU-Xray) [27], also known as the OpenI dataset, was released in 2016. This dataset was sourced from two large hospital systems within the Indiana Network for Patient Care database. It comprises 7,470 chest radiographs (including both frontal and

lateral views) and 3,955 corresponding narrative reports from 3,955 patients. Each report includes two primary sections: findings, which provide a detailed natural language description of the significant aspects in the image, and impression, which offer a concise summary of the most immediately relevant findings. The 3,955 studies are divided into 2,470 abnormal cases and 1,485 normal cases. Disease labels were extracted from the reports either manually or automatically using Medical Subject Headings (MeSH) [32], Radiology Lexicon (RadLex) [63], and the Medical Text Indexer (MTI) [90]. The ten most frequent disease tags are cardiomegaly, pulmonary atelectasis, calcified granuloma, tortuous aorta, hypoinflated lung, lung base opacity, pleural effusion, lung hyperinflation, lung cicatrix, and lung calcinosis. Since the dataset does not have an official split, the common practice is to randomly divide it into training, validation, and test sets in a 7:1:2 ratio.

MIMIC-CXR: The Medical Information Mart for Intensive Care Chest X-ray (MIMIC-CXR) [56] is the largest public medical image-report dataset. It includes imaging studies from 65,379 patients from the Beth Israel Deaconess Medical Center Emergency Department, collected between 2011 and 2016. The dataset includes 377,110 chest radiographs (including frontal and lateral views) and 227,835 corresponding reports, most of which include findings and impression sections. Each report is associated with to one or more images. On average, 3.5 reports from different time periods are collected for each patient, providing longitudinal data that allows researchers to reference previous images. The dataset is officially split into training, validation, and test sets, which improves reproducibility. Specifically, the training set contains 368,960 images and 222,758 reports, the validation set contains 2,991 images and 1,808 reports, and the test set contains 5,159 images and 3,269 reports.

The images were originally stored in DICOM format, but a JPEG version (MIMIC-CXR-JPG) [57] was also created to reduce storage size. In addition, 14 structured disease labels were extracted from the reports using NegBio [101] and Chexpert [49], including atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pneumonia, pneumothorax, pleural other, support devices, and no finding.

Due to the large volume and the diversity of diseases in the MIMIC-CXR dataset, two derived datasets were created. MIMIC-ABM [94] is a subset that contains only abnormal studies with at least one abnormal finding. This subset, consisting of 38,551 pairs (26,946 for training, 3,801 for validation, and 7,804 for testing), addresses data bias caused by the prevalence of normal studies in the original dataset, allowing models to learn abnormal patterns more effectively. Another derived dataset, MIMIC-PRO [104], eliminates all temporal information within the reports. Training with this dataset helps mitigate the generation of hallucinations about non-existent priors by the model.

5.2 Other Datasets

In addition to the benchmark dataset of chest radiographs with English reports, there are several public datasets containing non-English reports and non-radiographic images. These datasets expand the range of languages and imaging modalities, catering to the specific needs of different research communities.

Padchest: The Pathology Detection in Chest Radiographs (Padchest) [15] contains imaging studies of 67,625 patients collected from 2009 to 2017 at Hospital San Juan, Spain. It contains 160,868 chest radiographs and 109,931 reports. The radiographs include six views: postero-anterior (PA), lateral, AP-horizontal, AP-vertical, costal, and pediatric. All reports are written in Spanish, and each report potentially corresponds to one or multiple views.

CX-CHR: CX-CHR [76] contains 45,598 chest radiographs and 33,236 reports written in Chinese. Each report corresponds to one or multiple views, including PA and lateral views. The reports include findings and impression sections, as well as labels for 20 common chest diseases. Although this dataset is internally proprietary, researchers can apply for academic use after signing a confidentiality agreement, and it has been used in Wang *et al.* [120] and Li *et al.* [73].

COV-CTR: The COVID-19 CT Report dataset (COV-CTR) [73] contains 728 lung CT scans and corresponding reports. The images are from the public COVID-CT dataset [143], and the reports are written in Chinese by three radiologists from the First Affiliated Hospital of Harbin Medical University. Of these studies, 349 are COVID-19 cases, and 379 are non-COVID-19 cases.

FFA-IR: The Fundus Fluorescein Angiography Images and Reports (FFA-IR) [70] was collected from patients at the Zhongshan Ophthalmic Center of Sun Yat-Sen University in Guangzhou, China, between November 2016 and December 2019. The dataset comprises 1,048,584 FFA images and 10,790 reports, encompassing 46 categories of retinal lesions. Each report is bilingual, with both Chinese and English versions available. Approximately 5% of the cases are healthy, while the remaining cases present various retinal conditions. The dataset is divided into official splits: 8,016 cases for training, 1,069 cases for validation, and 1,604 cases for testing, facilitating future model performance comparisons. Each case includes not only the report and FFA images but also explainable annotations to enhance the interpretability of AMRG models. Specifically, ophthalmologists labeled the lesion locations and categories in the images with rectangular boxes based on the size, location, and stage of the lesions described in the report.

EndoVis-18: The EndoVis-18 dataset originates from the MICCAI Robotic Scene Segmentation of Endoscopic Vision Challenge 2018 [6]. It comprises 1,560 endoscopic surgical images, each annotated by experienced surgeons [135, 136], and corresponding scene graphs are generated by Islam

et al. [50]. The dataset includes a total of nine objects, featuring one type of tissue (kidney) and eight different surgical instruments. Additionally, there are 11 types of interactions between the surgical instruments and tissue, such as manipulation, grasping, and cutting. Following the methodology of previous studies, 1,124 images along with their captions and scene graphs are used as the training set, while the remaining images serve as the test set.

TORS: The TORS dataset was collected from transoral robotic surgery [135, 136] and consists of 335 surgical images, also annotated by experienced surgeons with associated scene graphs. It includes five types of objects: tissue, clip applier, suction, spatulated monopolar cautery, and Maryland dissector, with eight types of semantic interactions such as clipping, suturing, and grasping.

6 Evaluation Metrics

This section discusses various evaluation metrics used to evaluate the quality of generated reports, including (i) NLP metrics (Section 6.1), (ii) clinical efficacy (CE) metrics (Section 6.2), and (iii) human evaluation (Section 6.3). NLP metrics measure the word overlap between the generated report and the reference report, while CE metrics evaluate the clinical accuracy of the generated reports by focusing on specific disease labels. Human evaluation involves inviting radiologists to assess the generated reports to ensure reliability.

6.1 NLP metrics

NLP metrics were originally designed for natural language tasks and are employed in AMRG tasks to measure the quality of generated reports. Commonly used NLP metrics include BLEU, METEOR, ROUGE-L, and CIDEr, which are described in detail in the following sections.

BLEU: Bilingual Evaluation Understudy (BLEU) [100] was originally designed for machine translation. It measures the correspondence between a candidate (generated) sequence and a reference (ground-truth) sequence. A higher BLEU score indicates a closer match to the reference sequence.

BLEU calculates precision (p_n) for n-grams (subsequences of n words) and includes a brevity penalty (BP) for overly short sentences. BLEU is computed as follows:

$$\begin{aligned} p_n &= \frac{\text{Number of n-grams in candidate that match reference}}{\text{Total number of n-grams in candidate}} \\ BP &= \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}, \quad \text{BLEU} = BP \cdot \exp(\sum_{n=1}^{N} w_n \log p_n), \end{aligned} \tag{19}$$

where c and r are the lengths of the candidate and reference sequence, respectively, and w_n is the weight for n-grams (typically $w_n = \frac{1}{N}$).

METEOR: The Metric for Evaluation of Translation with Explicit Ordering (METEOR) [8] is a precision and recall-based measure that improves BLEU-1 by considering synonyms, stemming, and word order. It expands uni-gram matching to include exact matches, stemming matches, and synonym matches. METEOR aligns more closely with human judgment when comparing candidate and reference sequences.

Precision (P) and recall (R) are computed as follows:

$$P = \frac{m}{\text{length of candidate}}, \quad R = \frac{m}{\text{length of reference}}, \tag{20}$$

where m is the number of matches (including exact, stemmed, and synonym matches) between candidate and reference uni-grams.

The harmonic mean (F_{mean}) of precision and recall is:

$$F_{mean} = \frac{10PR}{R + 9P} \tag{21}$$

METEOR also includes a fragmentation penalty (Pen) to penalize candidate sequences with poor word order:

$$Pen = 0.5 \cdot \frac{\#\text{chunks}}{m}, \tag{22}$$

where #chunks are the number of groups of matched words in the same order as the reference. The final METEOR score is formulated as:

$$\text{METEOR} = F_{mean} \cdot (1 - Pen) \tag{23}$$

ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (ROUGE-L) [78] is a metric originally designed for automatic summarization. It measures the longest common subsequence (LCS) between the candidate (X) and reference (Y) sequences. A higher ROUGE-L score indicates better quality in terms of the structure and important content of the reference sequence.

Precision (P) and recall (R) are computed by:

$$P = \frac{LCS(X,Y)}{\text{length of candidate}}, \quad R = \frac{LCS(X,Y)}{\text{length of reference}}, \tag{24}$$

where $LCS(X,Y)$ denotes the length of the LCS of sequences X and Y.

ROUGE-L is formulated as:

$$\text{ROUGE-L} = \frac{(1+\beta^2)PR}{R + \beta^2 P}, \tag{25}$$

where β is a hyper-parameter that determines the relative importance of precision and recall. It is usually set to a large number to emphasize the recall score.

CIDEr: Consensus-based Image Description Evaluation (CIDEr) [119] was designed for image captioning. It measures the cosine similarity between a generated caption and a set of reference captions using the term frequency-inverse document frequency (TF-IDF) weighted n-grams. TF-IDF assigns higher weights to significant words, so a high CIDEr score indicates substantial coverage of these important words. TF quantifies how frequently an n-gram appears in a caption, while IDF measures how common an n-gram appears across all reference captions. The TF and IDF are calculated as follows:

$$\mathrm{TF}_{ij} = \frac{\mathrm{Count}_{ij}}{\sum_k \mathrm{Count}_{ik}}, \quad \mathrm{IDF}_i = \log \frac{N}{\sum_j \min(1, \mathrm{Count}_{ij})}, \tag{26}$$

where Count_{ij} is the count of the n-gram i in caption j, $\sum_k \mathrm{Count}_{ik}$ is the total count of all n-grams in caption j, N is the total number of reference captions, and $\sum_j \min(1, \mathrm{Count}_{ij})$ is the number of reference captions containing the n-gram i.

The TF-IDF weighting for an n-gram i in caption j is calculated as:

$$\text{TF-IDF}_{ij} = \mathrm{TF}_{ij} \cdot \mathrm{IDF}_i \tag{27}$$

Denote TF-IDF vector for n-grams of length n and caption j as $v_{j,n}$. The CIDEr_n score for n-grams of length n is computed by averaging the cosine similarity between the candidate and reference captions over all reference captions:

$$\mathrm{CIDEr}_n(c, R) = \frac{1}{|R|} \sum_{r \in R} S(v_{c,n}, v_{r,n}), \tag{28}$$

where R is the set of reference captions, $|R|$ is the number of reference captions, S represents cosine similarity, and $v_{c,n}$ and $v_{r,n}$ are TF-IDF vectors for the candidate and reference captions for n-grams of size n.

The final CIDEr score is a weighted average of the CIDEr_n scores for different n-gram lengths:

$$\mathrm{CIDEr}(c, R) = \frac{1}{N} \sum_{n=1}^{N} w_n \cdot \mathrm{CIDEr}_n(c, R), \tag{29}$$

where N is the maximum n-gram length (typically 4), and w_n is the weight for n-grams of length n (typically $w_n = \frac{1}{N}$).

6.2 Clinical Efficacy

NLP metrics primarily measure the word overlap between the generated report and the reference report. However, in the medical field, semantic similarity

and factual consistency between the generated report and the reference report are more important. For example, the NLP scores for the sentences "The heart is within normal size and contour" and "No cardiomegaly observed" are zero. However, in medical reports, these two sentences convey the same information. Therefore, many radiographic report generation studies supplement NLP metrics with clinical efficacy (CE) metrics [46, 85, 81, 129, 141, 86, 152, 30, 23, 139, 89, 22, 40, 102, 44, 80, 122]. Specifically, they use CheXpert [49] to extract labels from the generated and reference reports, focusing on 12 possible chest diseases. The label-based precision, recall, and F1 score are then calculated as CE metrics. This approach allows CE metrics to assess whether the generated report and the reference report contain the same diseases.

However, CE metrics are currently limited to the evaluation of English chest radiography reports. There are no tools similar to CheXpert for extracting disease labels from text for other body parts, modalities, and languages, which presents an opportunity for future research.

6.3 Human Evaluation

However, both NLP metrics and CE metrics sometimes are unreliable for evaluating medical reports, and CE metrics are limited to pre-trained disease categories. Some researchers suggest introducing human evaluation to comprehensively assess the quality of generated reports [89, 102, 73, 70, 155, 82]. Specifically, reports generated by multiple candidate models are mixed together, and multiple board-certified radiologists compare the candidate reports with the reference reports to avoid personal bias. The radiologists select the generated reports that are most similar to the reference reports based on fluency, factual consistency, and overall quality. While human evaluation is the most reliable evaluation method, it is also the most expensive and impractical for large-scale evaluations.

7 Performance Comparisons

Table 1 shows the results of SOTA radiographic reporting methods published between 2021 and 2024 on benchmark radiography datasets. By comparing their performance and techniques, we identified six techniques that effectively improve NLP and CE metrics: (i) human-computer interaction, (ii) reinforcement learning, (iii) detection and segmentation, (iv) disease classification, (v) traceback mechanism, and (vi) local alignment. In the following sections, BLEU-4, METEOR, ROUGE-L, and CIDEr in NLP metrics, as well as precision, recall, and F1 score in CE metrics, are simplified to BL-4, MTR, RG-L, CD, P, R, and F, respectively.

Human-computer interaction technology achieves the highest scores in both NLP and CE. Specifically, the inclusion of doctors' notes significantly enhances the quality of the generated reports. Nguyen *et al.* [93] achieved the best NLP scores on both MIMIC-CXR (BL-4: 0.224, MTR: 0.222, and RG-L: 0.390) and IU-Xray (BL-4: 0.235, MTR: 0.219, and RG-L: 0.436) by incorporating clinical documents. These clinical documents, which may include patients' clinical histories or doctors' notes, guide the model to focus on specific areas of the image and relevant diseases. For instance, if the doctor's note mentions "shows cough and shortness of breath symptoms", the model will focus on the lung area and consider pneumonia. In addition to human-computer interaction, the model involves disease classification, memory retrieval and traceback mechanism, as illustrated in Figure 9. Similarly, Liu *et al.* [86] introduced disease labels provided by radiologists, achieving the highest CE scores (P: 0.855, R: 0.730, and F: 0.773). Their model offers two options: automatic disease classification based on the input image or radiologist-provided potential disease labels. The latter results in markedly higher clinical efficacy in generated reports. Thus, incorporating human guidance into the model effectively improves the quality of the generated reports.

For methods without human interaction, reinforcement learning (Section 3.4.2) is most effective. Xu *et al.* [137] used BL-4, MTR, and CD as rewards and achieved the second-highest NLP scores (BL-4: 0.192, MTR: 0.207, and RG-L: 0.380) on the MIMIC-CXR dataset. Likewise, the factual completeness and consistency reward designed by Miura *et al.* [89] resulted in the second-highest CE scores (P: 0.503, R: 0.651, and F: 0.567) on MIMIC-CXR dataset and highest CD scores (CD: 0.509 and 1.034) on both datasets. In addition, using CD alone as a reward can also lead to relatively high NLP and CE scores on benchmark datasets [131, 80, 130]. Therefore, incorporating reinforcement learning into AMRG models is a straightforward but effective strategy.

Another effective technique involves leveraging pre-trained detection or segmentation networks (Section 3.2.2) to enhance AMRG models by focusing on meaningful anatomical regions. For instance, Tanida *et al.* [116] integrated a detection network with binary classifiers, as illustrated in Figure 10, enabling the model to concentrate on critical regions and achieving the second-highest CD score of 0.495 on the MIMIC-CXR dataset. Similarly, Zhao *et al.* [158] employed a segmentation network, securing the second-highest NLP scores (BL-4: 0.221 and RG-L: 0.433) on the IU-Xray dataset.

In addition, integrating multi-label disease classification (Section 3.2.1) yields considerable improvements in CE scores [54, 44, 141, 43, 85, 46], ensuring that the diseases identified in the generated reports are consistent with those in the input images. For implementation, the image encoder can use disease classification as a pre-training task, or the classification can be employed as a joint learning task during training. Notably, incorporating the classification results as additional information into the text decoder can yield higher CE

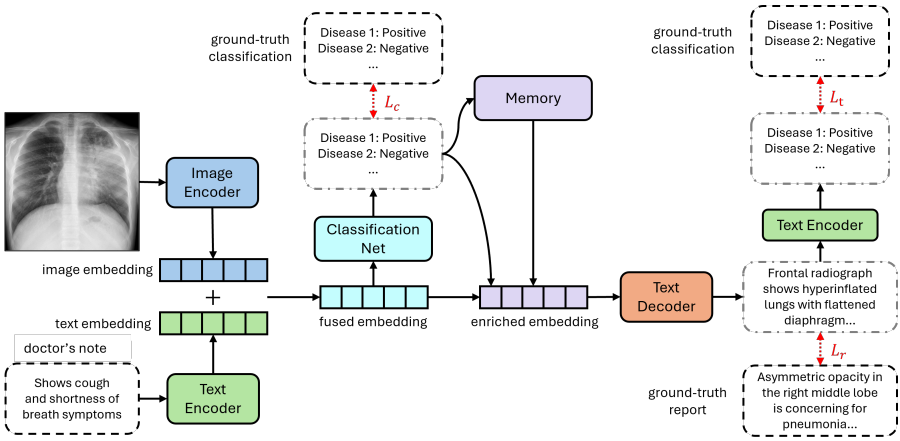

Figure 9: This diagram illustrates the architecture of the model developed by Nguyen *et al.* [93], which involves human-computer interaction, disease classification, memory retrieval, and a traceback mechanism. The model begins by encoding a chest radiograph and the doctor's note into image and text embeddings. These embeddings are combined to form a fused embedding, which is then processed by a classification network to predict the patient's diseases. The predicted diseases guide a search through stored memory to retrieve relevant information. The fused embedding, predicted diseases, and retrieved memory are integrated to create an enriched embedding, which is subsequently decoded into a report. This generated report is further classified using a text encoder-based classifier to verify whether the diseases identified in the report align with the diseases indicated by the image. Model optimization is driven by three loss functions: the classification loss ($\mathcal{L}_c$) between the predicted and ground truth (GT) diseases, the report loss ($\mathcal{L}_r$) between the generated and GT reports, and the traceback loss ($\mathcal{L}_t$) between the diseases in the generated report and the GT diseases. The total loss is $\mathcal{L}_{total} = \mathcal{L}_c + \mathcal{L}_r + \mathcal{L}_t$.

metrics. For example, Li *et al.* [54] reported CE scores of P: 0.501, R: 0.509, and F: 0.476, while Hou *et al.* [44] achieved CE scores of P: 0.416, R: 0.418, and F: 0.385.

For the IU-Xray dataset, both the traceback mechanism (Section 3.4.1) [65, 131, 145] and local alignment (Section 3.1.2) [125, 80, 145] demonstrate notable efficacy. The traceback mechanism involves making the generated report similar to the reference report at the feature level, while local alignment aligns sentences or words with image patches. In particular, Li *et al.* [65] employed a traceback mechanism to achieve notable NLP scores (BL-4: 0.215,and RG-L: 0.415), indicating that the generated reports are similar to the reference reports in terms of 4-gram and the longest sequence. Ye *et al.* [145] combined traceback with local alignment and obtained the highest MTR score of 0.233 and a relative high CD score of 0.469, indicating that the generated reports have a high overlap with reference reports in terms of keywords. However, these techniques did not yield similarly excellent results on the MIMIC-CXR

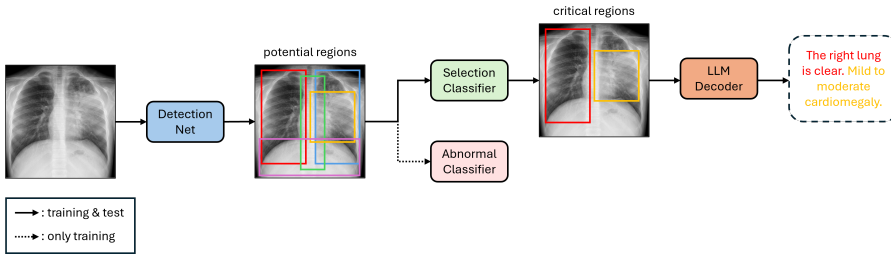

Figure 10: This diagram illustrates the architecture of the model developed by Tanida *et al.* [116], which integrates detection, disease classification, and a large language model (LLM). The detection network extracts visual features from 29 potential anatomical regions in chest radiographs. These features are then processed by an abnormal classifier and a selection classifier. The abnormal classifier determines whether a region contains a lesion, encoding strong abnormal information into the features. The selection classifier identifies regions critical for report generation, ensuring that only the visual features from these critical regions are passed to the decoder. The model's decoder, which incorporates a pre-trained LLM, injects the features of the selected critical regions through pseudo self-attention, generating a sentence for each region.

dataset, indicating that the traceback mechanism and local alignment still have limitations when applied to more complex and variable datasets.

8 Future Directions

Finally, we highlight unresolved issues in the current methods that present opportunities for future research in the AMRG field.

Multimodal Learning: The primary challenge in the AMRG field is bridging the modality gap between images and text. The CLIP model [103], which utilizes natural language as supervision for contrastive learning, has significantly advanced this area. However, from the current results (Table 1), the performance of SOTA AMRG models remains limited, with CIDEr scores much lower than those of SOTA general image captioning models [67, 68]. This discrepancy underscores the inadequacy of existing multimodal learning methods in fully supporting report generation models, particularly when dealing with medical images containing subtle differences.

One promising direction involves implementing local alignment techniques that associate specific image regions with textual entities, enabling the model to learn more fine-grained details. Although current local alignment methods have shown improvements on the IU-Xray dataset, they have not yielded significant benefits on more complex datasets such as MIMIC-CXR [125, 80, 145]. This indicates that current fine-grained alignment methods are still insufficient for the medical domain. Therefore, future research should focus

on advancing these methods to capture the nuanced and subtle features of medical images more effectively.

Unsupervised/Semi-Supervised Learning: Another major factor limiting the AMRG models is the relatively small size of paired medical datasets. For instance, the largest medical dataset, MIMIC-CXR [56], contains only 0.22 million pairs, whereas general image captioning datasets like Conceptual [18] contain 12 million pairs. The high cost of creating image-text paired medical datasets makes expanding them to a similar scale as general datasets impractical.

One potential solution is to employ unsupervised and semi-supervised learning to expand the available data. Some researchers have used image classification and text reconstruction to train image encoder and text decoder separately, allowing the model to learn valuable patterns and representations from unpaired data [40, 85, 132]. This approach enables the use of image-only and text-only medical data to augment the training set. However, two main limitations persist: the need for disease labels for images and the low precision of current methods. Image classification for training encoder requires images with disease labels, which is also labor-intensive. Moreover, as shown in Table 1, the performance of unsupervised [40, 85] and semi-supervised [152] methods is currently lower than that of supervised methods. Future research should focus on eliminating the need for image labels and improving the performance of unsupervised and semi-supervised methods. By using larger datasets than those used in supervised methods, their accuracy could ultimately surpass that of supervised methods.

Human-Computer Interaction: Given the limitations of current methods in terms of accuracy, human-computer interaction systems represent a viable avenue for further development. In such systems, physicians can provide prompts to guide the model in generating descriptive reports [93, 86], or the model can generate draft reports that physicians subsequently modify. Integrating the report generation model into the clinical diagnosis process can reduce repetitive tasks for physicians, allowing them to focus on diagnosing complex diseases. With physicians' supervision, the issue of low accuracy in the generated reports becomes less of an obstacle to clinical application, as physicians can refine the generated text.

Another approach to human-computer interaction involves incorporating physicians' feedback into the model's iteration process. This approach offers a more precise and targeted supervision signal than the loss function. By continuously optimizing the report generation model based on physicians' feedback during daily use, the model can become more adept at addressing specific diseases.

Interpretability: Another challenge to the clinical application of the AMRG models is the opacity of their decision-making process. Providing visual and textual explanations can assist physicians in understanding the

rationale behind specific diagnostic recommendations. A common approach is to utilize back-propagated gradients to highlight pertinent regions within the image [109]. However, given that the output of the AMRG models is lengthy text, this approach is unsuitable. To enhance interpretability, Chen *et al.* [20] attempted to show rectangular boxes on the image to indicate areas where the model believed lesions occurred and corresponded to specific generated sentences. Unfortunately, their model produced many overlapping boxes, which did not clearly explain the decision-making process.

A future direction for improving interpretability involves refining visual explanation techniques to highlight critical regions more precisely. Additionally, since descriptions of diseases in the report are more critical than those of normal conditions, combining visual explanations with textual ones would be beneficial. This approach can highlight key words or sentences in the generated text and the corresponding regions in the input image. Such a combination of visual and textual interpretation is more suitable for the AMRG domain.

Evaluation Metrics: Developing more accurate evaluation metrics to assess the accuracy of generated reports is also a critical need in current research. Current NLP evaluation metrics primarily measure word similarity between generated and reference texts, but they fail to capture the clinical accuracy of reports. Meanwhile, CE metrics are limited to chest radiographic reports and fixed disease categories. While some studies introduce human evaluation, its high cost and lack of standardized criteria hinder large-scale implementation.

The AMRG field needs specialized evaluation metrics or methods that can assess the correctness of medical terminology and the accuracy of diagnoses. These metrics should be also applicable to various image modalities and diseases, as well as handle the inherent variability in medical diagnoses. Additionally, such metrics should be scalable, cost-effective, and standardized to enable consistent comparisons of model performance, similar to existing NLP metrics.

9 Conclusion

In conclusion, the field of AMRG has made significant strides in recent years, addressing critical challenges and enhancing the efficiency and accuracy of medical diagnoses. Our comprehensive review of AMRG methods from 2021 to 2024 highlights fourteen solutions for the four primary challenges: modality gap, visual deviations, text complexity, and dataset limitations. We also present AMRG applications across various imaging modalities, including radiography, CT scans, MRI, ultrasound, FFA, endoscopic imaging, surgical scenes, and WSI. In addition, our review underscores the importance of publicly available datasets and robust evaluation metrics in advancing AMRG research. Based

on their performance on benchmark datasets, we identify six solutions that can significantly improve evaluation metrics.

Despite these advancements, the field continues to face ongoing challenges. Future research should focus on developing more effective multimodal learning algorithms, enhancing human-computer interaction, expanding available datasets, improving model interpretability, and refining evaluation metrics to ensure greater accuracy.

References

[1] W. Abdelrahman and A. Abdelmageed, "Medical record keeping: clarity, accuracy, and timeliness are essential", *BMJ: British Medical Journal*, 348, 2014, f7716.

[2] B. Abhisheka, S. K. Biswas, B. Purkayastha, D. Das, and A. Escargueil, "Recent trend in medical imaging modalities and their applications in disease diagnosis: a review", *Multimedia Tools and Applications*, 2023, 1–36.

[3] N. Aksoy, N. Ravikumar, and A. F. Frangi, "Radiology report generation using transformers conditioned with non-imaging data", in *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, Vol. 12469, SPIE, 2023, 146–54.

[4] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers", *Informatics in Medicine Unlocked*, 24, 2021, 100557.

[5] A. Aljuaid and M. Anwar, "Survey of supervised learning for medical image processing", *SN Computer Science*, 3(4), 2022, 292.

[6] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, *et al.*, "2018 robotic scene segmentation challenge", *arXiv preprint arXiv:2001.11190*, 2020.

[7] I. Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz, "Automatic caption generation for medical images", in *Proceedings of the 3rd International Conference on Smart City Applications*, 2018, 1–6.

[8] S. Banerjee and A. Lavie, "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*, 2005, 65–72.

[9] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 15016–27.

[10] D.-R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (MIC): a rapid review of literature", *Artificial intelligence review*, 56(5), 2023, 4019–76.

[11] L. Berlin, "Liability of interpreting too many radiographs", *American Journal of Roentgenology*, 175(1), 2000, 17–22.

[12] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?", *Insights into imaging*, 8, 2017, 171–82.

[13] A. Bria, C. Marrocco, and F. Tortorella, "Addressing class imbalance in deep learning for small lesion detection on medical images", *Computers in biology and medicine*, 120, 2020, 103735.

[14] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction", *Radiographics*, 35(6), 2015, 1668–76.

[15] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vaya, "Padchest: A large chest x-ray image dataset with multi-label annotated reports", *Medical image analysis*, 66, 2020, 101797.

[16] Y. Cao, L. Cui, L. Zhang, F. Yu, Z. Li, and Y. Xu, "MMTN: multi-modal memory transformer network for image-report consistent medical report generation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 1, 2023, 277–85.

[17] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, *et al.*, "A systematic review of natural language processing applied to radiology reports", *BMC medical informatics and decision making*, 21(1), 2021, 179.

[18] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, 3558–68.

[19] P. Chen, H. Li, C. Zhu, S. Zheng, and L. Yang, "MI-Gen: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images", *arXiv preprint arXiv:2311.16480*, 2023.

[20] W. Chen, X. Li, L. Shen, and Y. Yuan, "Fine-grained image-text alignment in medical imaging enables cyclic image-report generation", *arXiv preprint arXiv:2312.08078*, 2023.

[21] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, "Mapping medical image-text to a joint space via masked modeling", *Medical Image Analysis*, 91, 2024, 103018.

[22] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation", *arXiv preprint arXiv:2204.13258*, 2022.

[23] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer", *arXiv preprint arXiv:2010.16056*, 2020.

[24] P. Cheng, L. Lin, J. Lyu, Y. Huang, W. Luo, and X. Tang, "Prior: Prototype representation joint learning from medical images and reports", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 21361–71.

[25] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 10578–87.

[26] G. Dawidowicz, E. Hirsch, and A. Tal, "Limitr: Leveraging local information for medical image-text representation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 21165–73.

[27] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval", *Journal of the American Medical Informatics Association*, 23(2), 2016, 304–10.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, 248–55.

[29] T. Dyer, L. Dillard, M. Harrison, T. N. Morgan, R. Tappouni, Q. Malik, and S. Rasalingham, "Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm", *Clinical radiology*, 76(6), 2021, 473–e9.

[30] M. Endo, R. Krishnan, V. Krishna, A. Y. Ng, and P. Rajpurkar, "Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model", in *Machine Learning for Health*, PMLR, 2021, 209–19.

[31] R. Estopà and M. A. Montané, "Terminology in medical reports: Textual parameters and their lexical indicators that hinder patient understanding", *Terminology*, 26(2), 2020, 213–36.

[32] R. FB, "Medical subject headings.", *Bulletin of the Medical Library Association*, 51, 1963, 114–6.

[33] R. Fitzgerald, "Error in radiology", *Clinical radiology*, 56(12), 2001, 938–46.

[34] D. Gao, M. Kong, Y. Zhao, J. Huang, Z. Huang, K. Kuang, F. Wu, and Q. Zhu, "Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning", *Medical Image Analysis*, 91, 2024, 102982.

[35] E. V. Garcia, "Integrating artificial intelligence and natural language processing for computer-assisted reporting and report understanding in nuclear cardiology", *Journal of Nuclear Cardiology*, 30(3), 2023, 1180–90.

[36] M. Gatt, G. Spectre, O. Paltiel, N. Hiller, and R. Stalnikowicz, "Chest radiographs in the emergency department: is the radiologist really necessary?", *Postgraduate medical journal*, 79(930), 2003, 214–7.

[37] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction", *BMC bioinformatics*, 22, 2021, 1–20.

[38] Z. Han, B. Wei, X. Xi, B. Chen, Y. Yin, and S. Li, "Unifying neural learning and symbolic reasoning for spinal medical report generation", *Medical image analysis*, 67, 2021, 101872.

[39] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis", *Intelligent Medicine*, 3(1), 2023, 59–78.

[40] E. Hirsch, G. Dawidowicz, and A. Tal, "MedCycle: Unpaired Medical Report Generation via Cycle-Consistency", *arXiv preprint arXiv:2403.13444*, 2024.

[41] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration", *arXiv preprint arXiv:1904.09751*, 2019.

[42] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "Ratchet: Medical transformer for chest x-ray diagnosis and reporting", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, Springer, 2021, 293–303.

[43] W. Hou, Y. Cheng, K. Xu, W. Li, and J. Liu, "RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning", *arXiv preprint arXiv:2310.13864*, 2023.

[44] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "ORGAN: observation-guided radiology report generation via tree reasoning", *arXiv preprint arXiv:2306.06466*, 2023.

[45] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 3942–51.

[46] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 19809–18.

[47] S. Hussain, I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, and M. A. Sultan, "Modern diagnostic imaging technique applications and risk factors in the medical field: a review", *BioMed research international*, 2022(1), 2022, 5164970.

[48] B. M. Idowu and T. A. Okedere, "Diagnostic radiology in Nigeria: a country report", *Journal of Global Radiology*, 6(1), 2020.

[49] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, 2019, 590–7.

[50] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, Springer, 2020, 627–36.

[51] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports", *arXiv preprint arXiv:2106.14463*, 2021.

[52] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, and P. Rajpurkar, "Multimodal image-text matching improves retrieval-based chest x-ray report generation", in *Medical Imaging with Deep Learning*, PMLR, 2024, 978–90.

[53] Z. Ji, M. A. Shaikh, D. Moukheiber, S. N. Srihari, Y. Peng, and M. Gao, "Improving joint learning of chest x-ray and radiology report by word region alignment", in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, Springer, 2021, 110–9.

[54] H. Jin, H. Che, Y. Lin, and H. Chen, "Promptmrg: Diagnosis-driven prompts for medical report generation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 3, 2024, 2607–15.

[55] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports", *arXiv preprint arXiv:1711.08195*, 2017.

[56] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports", *Scientific data*, 6(1), 2019, 317.

[57] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs", *arXiv preprint arXiv:1901.07042*, 2019.

[58] N. Kaur, A. Mittal, and G. Singh, "Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey", *Multimedia Tools and Applications*, 81(10), 2022, 13409–39.

[59] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling", *arXiv preprint arXiv:1610.04325*, 2016.

[60] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", *arXiv preprint arXiv:1609.02907*, 2016.

[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 4015–26.

[62] E. A. Krupinski, K. S. Berbaum, R. T. Caldwell, K. M. Schartz, and J. Kim, "Long radiology workdays reduce detection and accommodation accuracy", *Journal of the American College of Radiology*, 7(9), 2010, 698–704.

[63] C. P. Langlotz, "RadLex: a new method for indexing online educational materials", 2006.

[64] K. H. Leung, S. P. Rowe, J. P. Leal, S. Ashrafinia, M. S. Sadaghiani, H. W. Chung, P. Dalaie, R. Tulbah, Y. Yin, R. VanDenBerg, *et al.*, "Deep learning and radiomics framework for PSMA-RADS classification of prostate cancer on PSMA PET", *EJNMMI research*, 12(1), 2022, 76.

[65] J. Li, S. Li, Y. Hu, and H. Tao, "A self-guided framework for radiology report generation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, 588–98.

[66] J. Li, T. Su, B. Zhao, F. Lv, Q. Wang, N. Navab, Y. Hu, and Z. Jiang, "Ultrasound Report Generation with Cross-Modality Feature Alignment via Unsupervised Guidance", *arXiv preprint arXiv:2406.00644*, 2024.

[67] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", in *International conference on machine learning*, PMLR, 2023, 19730–42.

[68] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", in *International conference on machine learning*, PMLR, 2022, 12888–900.

[69] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation", *Advances in neural information processing systems*, 34, 2021, 9694–705.

[70] M. Li, W. Cai, R. Liu, Y. Weng, X. Zhao, C. Wang, X. Chen, Z. Liu, C. Pan, M. Li, *et al.*, "Ffa-ir: Towards an explainable and reliable medical report generation benchmark", in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

[71] M. Li, W. Cai, K. Verspoor, S. Pan, X. Liang, and X. Chang, "Cross-modal clinical graph transformer for ophthalmic report generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 20656–65.

[72] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 3334–43.

[73] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation", *World Wide Web*, 26(1), 2023, 253–70.

[74] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm", *arXiv preprint arXiv:2110.05208*, 2021.

[75] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, and Y. Zou, "Unify, align and refine: Multi-level semantic alignment for radiology report generation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 2863–74.

[76] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation", *Advances in neural information processing systems*, 31, 2018.

[77] C. Lin, Z. Zhu, Y. Zhao, Y. Zhang, K. He, and Y. Zhao, "SGT++: Improved Scene Graph-guided Transformer for Surgical Report Generation", *IEEE Transactions on Medical Imaging*, 2023.

[78] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries", in *Text summarization branches out*, 2004, 74–81.

[79] S. W. Lindley, E. M. Gillies, and L. A. Hassell, "Communicating diagnostic uncertainty in surgical pathology reports: disparities between sender and receiver", *Pathology-Research and Practice*, 210(10), 2014, 628–33.

[80] A. Liu, Y. Guo, J.-h. Yong, and F. Xu, "Multi-grained Radiology Report Generation with Sentence-level Image-language Contrastive Learning", *IEEE Transactions on Medical Imaging*, 2024.

[81] C. Liu, Y. Tian, W. Chen, Y. Song, and Y. Zhang, "Bootstrapping Large Language Models for Radiology Report Generation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 17, 2024, 18635–43.

[82] F. Liu, S. Ge, Y. Zou, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation", *arXiv preprint arXiv:2206.14579*, 2022.

[83] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, 13753–62.

[84] F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation", *arXiv preprint arXiv:2106.06965*, 2021.

[85] F. Liu, C. You, X. Wu, S. Ge, X. Sun, *et al.*, "Auto-encoding knowledge graph for unsupervised medical report generation", *Advances in Neural Information Processing Systems*, 34, 2021, 16266–79.

[86] Z. Liu, Z. Zhu, S. Zheng, Y. Zhao, K. He, and Y. Zhao, "From Observation to Concept: A Flexible Multi-view Paradigm for Medical Report Generation", *IEEE Transactions on Multimedia*, 2023.

[87] Y. Lu, S. Hong, Y. Shah, and P. Xu, "Effectively fine-tune to improve large multimodal models for radiology report generation", *arXiv preprint arXiv:2312.01504*, 2023.

[88] S. Mayor, "Waiting times for x ray results in England are increasing, figures show", 2015.

[89] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation", *arXiv preprint arXiv:2010.10042*, 2020.

[90] J. G. Mork, A. Jimeno-Yepes, A. R. Aronson, *et al.*, "The NLM Medical Text Indexer System for Indexing Biomedical Literature.", *BioASQ@ CLEF*, 1, 2013.

[91] S. K. Mun, K. H. Wong, S.-C. B. Lo, Y. Li, and S. Bayarsaikhan, "Artificial intelligence for the future radiology diagnostic service", *Frontiers in molecular biosciences*, 7, 2021, 614258.

[92] I. Najdenkoska, X. Zhen, M. Worring, and L. Shao, "Uncertainty-aware report generation for chest X-rays by variational topic inference", *Medical Image Analysis*, 82, 2022, 102603.

[93] H. T. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng, "Automated generation of accurate\& fluent medical x-ray reports", *arXiv preprint arXiv:2108.12126*, 2021.

[94] J. Ni, C.-N. Hsu, A. Gentili, and J. McAuley, "Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays", *arXiv preprint arXiv:2010.02467*, 2020.

[95] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting", *Artificial intelligence in medicine*, 144, 2023, 102633.

[96] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding", *arXiv preprint arXiv:1807.03748*, 2018.

[97] R. Pan, R. Ran, W. Hu, W. Zhang, Q. Qin, and S. Cui, "S3-Net: A Self-Supervised dual-Stream Network for Radiology Report Generation", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[98] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 10971–80.

[99] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning", *BioMedical Engineering OnLine*, 22(1), 2023, 48.

[100] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation", in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, 311–8.

[101] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports", *AMIA Summits on Translational Science Proceedings*, 2018, 2018, 188.

[102] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation", in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, 448–58.

[103] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision", in *International conference on machine learning*, PMLR, 2021, 8748–63.

[104] V. Ramesh, N. A. Chi, and P. Rajpurkar, "Improving radiology report generation systems by removing hallucinated references to non-existent priors", in *Machine Learning for Health*, PMLR, 2022, 456–73.

[105] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college", *BMJ: British Medical Journal (Online)*, 359, 2017.

[106] A. B. Rosenkrantz, D. R. Hughes, and R. Duszak Jr, "The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets", *Radiology*, 279(1), 2016, 175–84.

[107] D. A. Rosman, J. J. Nshizirungu, E. Rudakemwa, C. Moshi, J. de Dieu Tuyisenge, E. Uwimana, and L. Kalisa, "Imaging in the land of 1000 hills: Rwanda radiology country report", *Journal of Global Radiology*, 1(1), 2015.

[108] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 815–23.

[109] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in *Proceedings of the IEEE international conference on computer vision*, 2017, 618–26.

[110] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, 2556–65.

[111] H.-C. Shin, L. Lu, and R. M. Summers, "Natural language processing for large-scale medical image analysis using deep learning", *Deep learning for medical image analysis*, 2017, 405–21.

[112] J. Si, H. Zhao, L. Huang, and Z. Tu, "Non-symmetrical Sibling-Stream Network with Adaptive Positional Encoding for Automatic Medical Report Generation", in *2023 5th International Conference on Intelligent Medicine and Image Processing (IMIP)*, IEEE, 2023, 97–103.

[113] P. Sloan, P. Clatworthy, E. Simpson, and M. Mirmehdi, "Automated Radiology Report Generation: A Review of Recent Advances", *IEEE Reviews in Biomedical Engineering*, 2024.

[114] A. Srinivasa Babu and M. L. Brooks, "The malpractice liability of radiology reports: minimizing the risk", *Radiographics*, 35(2), 2015, 547–54.

[115] S. Srinivasan, D. Francis, S. K. Mathivanan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "A hybrid deep CNN model for brain tumor image multi-classification", *BMC Medical Imaging*, 24(1), 2024, 21.

[116] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 7433–42.

[117] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, *et al.*, "Towards generalist biomedical AI", *NEJM AI*, 1(3), 2024, AIoa2300138.

[118] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, 30, 2017.

[119] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 4566–75.

[120] F. Wang, X. Liang, L. Xu, and L. Lin, "Unifying relational sentence generation and retrieval for medical image report composition", *IEEE transactions on cybernetics*, 52(6), 2020, 5015–25.

[121] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation", in *European Conference on Computer Vision*, Springer, 2022, 563–79.

[122] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "CAMANet: class activation map guided attention network for radiology report generation", *IEEE Journal of Biomedical and Health Informatics*, 2024.

[123] L. Wang, M. Ning, D. Lu, D. Wei, Y. Zheng, and J. Chen, "An inclusive task-aware framework for radiology report generation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, 568–77.

[124] R. Wang, X. Wang, Z. Xu, W. Xu, J. Chen, and T. Lukasiewicz, "MvCo-DoT: Multi-View Contrastive Domain Transfer Network for Medical Report Generation", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[125] S. Wang, B. Peng, Y. Liu, and Q. Peng, "Fine-grained medical vision-language representation learning for radiology report generation", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, 15949–56.

[126] Y. Wang, K. Wang, X. Liu, T. Gao, J. Zhang, and G. Wang, "Self adaptive global-local feature enhancement for radiology report generation", in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, 2275–9.

[127] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated radiographic report generation purely on transformer: A multicriteria supervised approach", *IEEE Transactions on Medical Imaging*, 41(10), 2022, 2803–13.

[128] Z. Wang, L. Liu, L. Wang, and L. Zhou, "Metransformer: Radiology report generation by transformer with multiple learnable expert tokens", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 11558–67.

[129] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2gengpt: Radiology report generation with frozen llms", *Meta-Radiology*, 1(3), 2023, 100033.

[130] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, 655–64.

[131] Z. Wang, L. Zhou, L. Wang, and X. Li, "A self-boosting framework for automated radiographic report generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 2433–42.

[132] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text", *arXiv preprint arXiv:2210.10163*, 2022.

[133] Y. Wen, L. Chen, L. Qiao, Y. Deng, S. Dai, J. Chen, and C. Zhou, "Symptom and pathology report generation for ophthalmic diseases in fundus images", in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2020, 349–56.

[134] X. Wu, J. Li, J. Wang, and Q. Qian, "Multimodal contrastive learning for radiology report generation", *Journal of Ambient Intelligence and Humanized Computing*, 14(8), 2023, 11185–94.

[135] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Class-incremental domain adaptation with smoothing and calibration for surgical report generation", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, Springer, 2021, 269–78.

[136] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Learning domain adaptation with model calibration for surgical report generation in robotic surgery", in *2021 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2021, 12350–6.

[137] Z. Xu, W. Xu, R. Wang, J. Chen, C. Qi, and T. Lukasiewicz, "Hybrid reinforced medical report generation with m-linear attention and repetition penalty", *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[138] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, Springer, 2018, 457–66.

[139] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, and C.-N. Hsu, "Weakly supervised contrastive learning for chest x-ray report generation", *arXiv preprint arXiv:2109.12242*, 2021.

[140] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed abnormality graph embedding for clinically accurate x-ray report generation", *IEEE Transactions on Medical Imaging*, 42(8), 2023, 2211–22.

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

[142] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge", *Medical image analysis*, 80, 2022, 102510.

[143] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19", *arXiv preprint arXiv:2003.13865*, 2020.

[144] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation", *IEEE Transactions on Multimedia*, 25, 2021, 167–78.

[145] S. Ye, M. Meng, M. Li, D. Feng, and J. Kim, "Dual-modal Dynamic Traceback Learning for Medical Report Generation", *arXiv preprint arXiv:2401.13267*, 2024.

[146] X. Yi, Y. Fu, R. Liu, H. Zhang, and R. Hua, "TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation", *IEEE Journal of Biomedical and Health Informatics*, 2024.

[147] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, 72–82.

[148] J. You, D. Li, M. Okumura, and K. Suzuki, "Jpg-jointly learn to align: Automated disease prediction and radiology report generation", in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, 5989–6001.

[149] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, 101–11.

[150] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, *et al.*, "Evaluating progress in automatic chest x-ray radiology report generation", *Patterns*, 4(9), 2023.

[151] J. Zhang, X. Shen, S. Wan, S. K. Goudos, J. Wu, M. Cheng, and W. Zhang, "A novel deep learning model for medical report generation by inter-intra information calibration", *IEEE Journal of Biomedical and Health Informatics*, 27(10), 2023, 5110–21.

[152] K. Zhang, H. Jiang, J. Zhang, Q. Huang, J. Fan, J. Yu, and W. Han, "Semi-supervised medical report generation via graph-guided hybrid feature consistency", *IEEE Transactions on Multimedia*, 26, 2023, 904–15.

[153] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 07, 2020, 12910–7.

[154] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text", in *Machine Learning for Healthcare Conference*, PMLR, 2022, 2–25.

[155] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. P. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports", *arXiv preprint arXiv:1911.02541*, 2019.

[156] Z. Zhang, B. Wang, W. Liang, Y. Li, X. Guo, G. Wang, S. Li, and G. Wang, "Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning", in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 1731–5.

[157] G. Zhao, Y. Yan, and Z. Zhao, "Normal-Abnormal Decoupling Memory for Medical Report Generation", in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, 1962–77.

[158] R. Zhao, X. Wang, H. Dai, P. Gao, and P. Li, "Medical Report Generation Based on Segment-Enhanced Contrastive Representation Learning", in *CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, 2023, 838–49.

[159] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models", *arXiv preprint arXiv:2304.10592*, 2023.

[160] Z. M. Ziegler, L. Melas-Kyriazi, S. Gehrmann, and A. M. Rush, "Encoder-agnostic adaptation for conditional language generation", *arXiv preprint arXiv:1908.06938*, 2019.