Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation

Shenshen Bu, Taiji Li, Yuedong Yang,* Zhiming Dai *
School of Computer Science and Engineering, Sun Yat-sen University, China
{bushsh, litj5}@mail2.sysu.edu.cn, {yangyd25, daizhim}@mail.sysu.edu.cn

Abstract

Automatic radiology report generation can provide substantial advantages to clinical physicians by effectively reducing their workload and improving efficiency. Despite the promising potential of current methods, challenges persist in effectively extracting and preventing degradation of prominent features, as well as enhancing attention on pivotal regions. In this paper, we propose an Instance-level Expert Knowledge and Aggregate Discriminative Attention framework (EKAGen[1]) for radiology report generation. We convert expert reports into an embedding space and generate comprehensive representations for each disease, which serve as Preliminary Knowledge Support (PKS). To prevent feature disruption, we select the representations in the embedding space with the smallest distances to PKS as Rectified Knowledge Support (RKS). Then, EKAGen diagnoses the diseases and retrieves knowledge from RKS, creating Instance-level Expert Knowledge (IEK) for each query image, boosting generation. Additionally, we introduce Aggregate Discriminative Attention Map (ADM), which uses weak supervision to create maps of discriminative regions that highlight pivotal regions. For training, we propose a Global Information Self-Distillation (GID) strategy, using an iteratively optimized model to distill global knowledge into EKAGen. Extensive experiments and analyses on IU X-Ray and MIMIC-CXR datasets demonstrate that EKAGen outperforms previous state-of-the-art methods.

1. Introduction

Radiology reports play a crucial role in the medical diagnosis and treatment process. However, interpreting radiology image can be extremely time-consuming, even for experienced radiologists. Hence, the automatic report generation [25, 29, 44, 47, 55] has emerged as a prominent research area within the medical imaging community. In recent years, significant strides have been made in the automatic generation of radiology reports, benefiting from the development of image captioning [2, 7, 14, 16, 21].

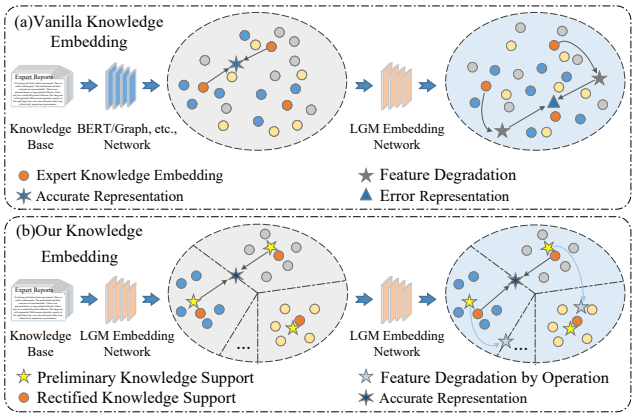

Figure 1. (a) displays the feature degradation caused by the distinction between vanilla methods' prior knowledge and LGM's embedding networks. (b) demonstrates our method that unifies the embedding network to prevent feature degradation and generates instance-level expert knowledge based on diseases.

In addition to the inherent challenges of natural image captioning, radiology report generation suffers three additional bottlenecks [27, 31, 47]. Firstly, radiology images often lack discriminative features, which results in a scarcity of reference information for the report generation models. Secondly, abnormal lesions in medical images may not always have obvious appearances, making them challenging even for experienced radiologists to identify. Thirdly, there is significant data deviation in these datasets due to the rarity of certain diseases, making it challenging to collect positive samples. In recent years, several methods, including the template retrieval structure [8, 31], memory driven network [5, 6], and knowledge aware module [27, 34, 49], were developed to address these challenges and have shown promising results in report generation. For instance, R2GenCMN [6] leveraged a shared memory to capture the alignment between images and texts. GSKET [49] proposed a multi-head attention mechanism that enhances

*Corresponding authors.
[1] https://github.com/hnjzbss/EKAGen

generation by combining visual features and knowledge.

However, as shown in Figure 1, (a) depicts that vanilla methods use separate embedding networks such as BERT [9, 48, 49] and Graph [15, 26, 54] to encode prior knowledge. This inconsistency between the embedding network for prior knowledge and the Language Generation Model (LGM) results in feature degradation, which causes prediction shifts when mapping word vectors back to the word space. As depicted in Figure 1 (b), our approach unifies the embedding network for prior knowledge and the Language Generation Model, and assigns instance-level prior knowledge to each instance based on the disease category, which differs to previous methods [15, 27, 49] incorporating general knowledge for every case, effectively resolving this issue. Additionally, previous methods [6, 18, 29, 31] fail to enhance attention on pivotal regions of the radiology image, thereby presenting certain limitations.

Our method is based on two implicit and valuable priors: (i) Diseases with the same label should have similar representations in the feature space, and vice versa; and (ii) Features from pivotal regions offer richer and more meaningful semantic references. Guided by these priors, we propose an Instance-level Expert Knowledge and Aggregate Discriminative Attention framework (**EKAGen**) for radiology report generation. We use an embedding network that shares parameters with LGM to map the expert report set into an embedding space as support set. From this, we calculate the Preliminary Knowledge Support (**PKS**) for each disease by taking the mean of its support set in the embedding space. Then, we compute the distances between the representations of each disease in PKS and all representations of corresponding disease in embedding space and select the representations with the smallest distances as the Rectified Knowledge Support (**RKS**). For a query image, we utilize a DiagnosisBot to identify diseases and retrieve their corresponding representations from RKS. These representations are then combined to create Instance-level Expert Knowledge (**IEK**), which facilitates report generation during the decoding process. In addition, identifying abnormal regions in images is crucial for report generation. However, obtaining pixel-level annotations for these regions is laborious and expensive. To tackle this challenge, we propose the Aggregate Discriminative Attention Map (ADM), which can generate discriminative regions of multiple diseases in a radiology image and integrate them to enhance attention on pivotal regions, thereby boosting generation.

Our contributions can be summarized as follows:

- We develop comprehensive embedding representations for each disease. By diagnosing the health conditions of different cases, we create Instance-level Expert Knowledge to provide our EKAGen with expert insights during the decoding process, addressing the issues of complex knowledge extraction and prominent feature degradation.
- To highlight pivotal regions, we employ weak supervision to generate activation maps, which are then used to create Aggregate Discriminative Attention Map (ADM). The ADM prioritizes key regions for each disease, reducing background noise and enhancing generation.
- To prevent potential feature erosion and provide more soft supervision, we propose a Global Information Self-Distillation (GID) strategy which utilizes an iteratively optimized model to distill global knowledge into our EKAGen, enhancing generation without additional labels.

2. Related Work

2.1. Image Captioning

Traditional image captioning methods typically rely on curated image-caption pairs to train an encoder-decoder model for generating text descriptions from input images. Early approaches [10, 12, 43] in this field employed a CNN-based encoder to extract visual features and an RNN/LSTM-based decoder for generating output sentences. To enhance visual comprehension, certain methods [2, 7, 17, 39] incorporated an object detector to identify and extract salient image regions. To foster greater interaction between the two modalities, attention mechanisms [7, 35, 36, 53] and graph neural networks [50, 51] gained widespread adoption. Recently, several impressive large-scale visual-language pre-training models [1, 16, 23, 24, 52] emerged, demonstrating outstanding performance in image captioning task. Natural scene image description focuses on generating concise sentences, while medical report generation requires detailed descriptions of medical images. Therefore, these approaches may not be suitable for the field of medical report generation.

2.2. Medical Report Generation

Medical report generation, as an extension of image captioning, presents greater challenges with higher requirements for text description length and accuracy. Extensive research [4, 22, 42, 46] has yielded significant advancements in this task. PPKED [32] presented an approach that explored and distilled posterior and prior knowledge in radiology to mitigate data bias issues in report generation. Clinical-BERT [48] proposed a visual-language pre-training model that learned medical domain knowledge to enhance the performance of report generation. DCL [26] extracted specific knowledge from retrieved reports to modify the graph structure and integrated image features with updated graphs to improve textual representation. ME-Transformer [47] introduced learnable "expert" tokens in the encoder and decoder, allowing them to interact with vision tokens and enabling the model to focus on different regions, while an orthogonal loss minimizes overlap to capture distinct information. However, these methods have drawbacks such as complex prior knowledge extraction,

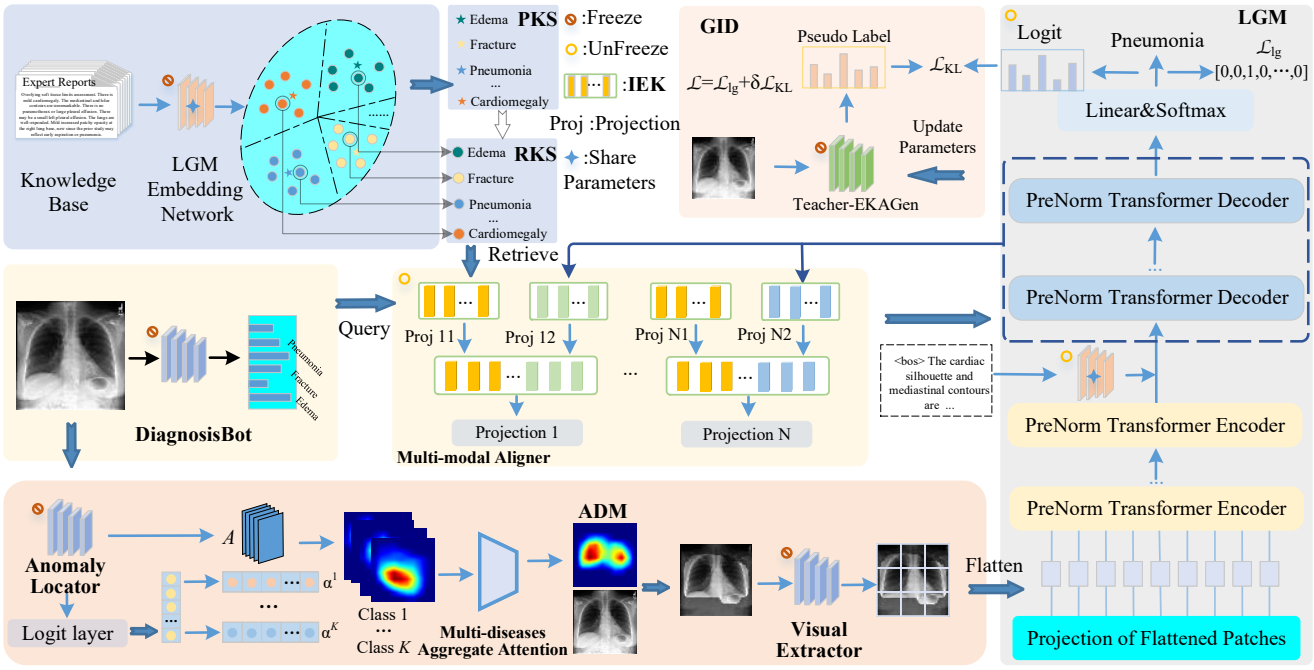

Figure 2. EKAGen consists of four components: Instance-level Expert Knowledge (IEK), Aggregate Discriminative Attention Map (ADM), Global Information Self-Distillation (GID), and Language Generation Model (LGM). EKAGen utilizes IEK to address the problem of feature degradation, employs ADM to prioritize pivotal regions, and incorporates GID to distill global knowledge.

prominent feature degradation, and inability to enhance attention on pivotal regions, resulting in certain limitations.

3. Method

In this section, we present a comprehensive analysis of the generation process for Instance-level Expert Knowledge and Aggregate Discriminative Attention Map. Additionally, we describe the structure of our Language Generation Model and elaborate on the Global Information Self-Distillation strategy. Figure 2 shows our EKAGen structure.

3.1. Instance-level Expert Knowledge

In order to address the problem of feature degradation of prior knowledge during the embedding process and the sparsity of features in radiology images, we propose the Instance-level Expert Knowledge (IEK) to boost generation.

Preliminary Theory of Knowledge Support Given an expert report set $\mathcal{X} = \{x_1, ..., x_N\}$ with C types of diseases, we compute the average similarity between x_i belonging to class c with the remaining cases within that class, and take the average as the panoptic score for x_i. The case with the highest panoptic score is selected as the Knowledge Support. This process formulated as follows:

$$\mathcal{K}_c = \underset{\mathcal{F}(x_i)}{\arg\max} \frac{1}{|\mathcal{X}_c|} \sum_{x_j \in \mathcal{X}_c} S(\mathcal{F}(x_i), \mathcal{F}(x_j)), x_i \in \mathcal{X}_c \quad (1)$$

where $S(\cdot)$ is similarity metric, and $\mathcal{F}(\cdot)$ is the embedding network that shares parameters with LGM, used to map reports to embedding space. $\mathcal{X}_c \subseteq \mathcal{X}$ is all cases in class c.

Preliminary Knowledge Support (PKS) The computational complexity of Equation 1 is $O(\sum_c N_c^2)$, which incurs high time cost. Inspired by the prototype [28, 41] paradigm, we first calculate the mean of all cases in class c as the PKS, serving as a compressed representation for each disease:

$$\mathcal{P}_c = \frac{1}{|\mathcal{X}_c|} \sum_{x_i \in \mathcal{X}_c} \mathcal{F}(x_i) \quad (2)$$

Rectified Knowledge Support (RKS) As shown in Figure 1 (b), Formula 2 directly operates on the embedding features, potentially causing disruption to the original word features and leading to catastrophic forgetting. To tackle this issue, we compute the cosine similarity between $\mathcal{P}_c$ and all cases in class c. The cases with the highest similarity are considered as the Rectified knowledge Support $\mathcal{K}$:

$$\mathcal{K}_c = \underset{\mathcal{F}(x_i)}{\arg\max} \frac{\mathcal{P}_c^T \mathcal{F}(x_i)}{||\mathcal{P}_c|| \cdot ||\mathcal{F}(x_i)||}, x_i \in \mathcal{X}_c \quad (3)$$

where $\mathcal{K}_c \in \mathbb{R}^{l_c \times d}$, l_c is the number of tokens and d is the embedding dimension. The combined computational complexities of Equation 2 and Equation 3 is $O(\sum_c N_c)$, significantly lower than that of Equation 1.

Expert Knowledge Navigator (EKN) In order to obtain the instance level expert knowledge, we utilize a multi-classification method as DiagnosisBot to detect diseases for

a given image I. By applying a threshold operation, we derive the category matrix $\mathcal{C}_i$ for the i-th disease:

$$logit = DiagnosisBot(I) \tag{4}$$

$$\mathcal{C}_i = \begin{cases} \mathbf{E}, & \text{if } \sigma(logit_i) > thre_i \\ \mathbf{0}, & otherwise \end{cases} \tag{5}$$

where $\mathbf{E} \in \mathbb{R}^{d\times d}$ represents the identity matrix and $\mathbf{0} \in \mathbb{R}^{d\times d}$ represents the zero matrix. We conduct matrix multiplication between the $\mathcal{K}_i$ and $\mathcal{C}_i$, followed by concatenation to generate Instance-level Expert Knowledge $\mathcal{K}^I$:

$$\mathcal{K}^I = concat(\mathcal{K}_1\mathcal{C}_1, ..., \mathcal{K}_i\mathcal{C}_i) \tag{6}$$

3.2. Aggregate Discriminative Attention Map

The radiology image differs from natural scene images in that it contains more noise and the discriminative areas are often blurry. Inspired by Grad CAM [40], we introduce the Aggregate Discriminative Attention Map (ADM), which leverages weak supervision signal to generate discriminative regions while attenuating the background.

Gradient-weighted Class Activation Mapping To obtain the class discrimination localization map $L^c_{GradCAM}$, Grad-CAM initially calculates the gradient of y^c with respect to the feature maps A. These gradients utilize GAP to calculate the weight α^c_k for k-th feature map and c-th class:

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{7}$$

where Z are the number of pixels, y^c is activation class score for the c-th class, and A^k_{ij} is activation of cell at spatial location i, j for k-th feature map. Then, a weighted combination is performed followed by $ReLU$ activation:

$$L^c_{GradCAM} = ReLU\sum_k \alpha^c_k A^k_{ij} \tag{8}$$

Multi-diseases Aggregate Attention We utilize Grad-CAM to generate activation maps highlighting the salient regions. Specifically, a multi-classification method is used as Anomaly Locator, which produces logit scores, with the scores exceeding the threshold considered valid. For image I with K valid classes, we combine the activation maps and apply thresholding to identify regions with strong signals:

$$M^I = \sum_{c=1}^{K} ReLU\sum_k \alpha^c_k A^k_{ij} \tag{9}$$

$$\mathcal{M}^I_{i,j} = \begin{cases} 1, & \text{if } M_{i,j} > \theta \\ 0, & otherwise \end{cases} \tag{10}$$

During visual analysis, we notice multiple gaps in $\mathcal{M}^I$. Therefore, we fill these gaps using morphological dilation to generate Aggregate Discriminative Attention Map $\mathcal{A}^I$:

$$\mathcal{A}^I_{ij} = \max_{(i',j')\in S} \mathcal{M}^I_{i+i',j+j'} \tag{11}$$

where S defines the neighborhood range, we can then use $\mathcal{A}^I$ as the foreground mask to diminish the background thereby boosting the discriminative regions in the image I:

$$Img_{aug} = \mathcal{A}^I \odot I + \gamma(1 - \mathcal{A}^I) \odot I \tag{12}$$

where γ represents the scaling factor for the background, and $\odot$ denotes the Hadamard Product operation.

3.3. Language Generation Model

Multi-modal Aligner We leverage linear layers to process the expert knowledge $\mathcal{K}^I$ for image I and the hidden features $\mathcal{H}^{i-1} \in \mathbb{R}^{n\times d}$ from the previous (i-1)-th decoder output. These features are then concatenated and projected to generate the multi-modal feature $\mathcal{M}^i \in \mathbb{R}^{n\times d}$:

$$\begin{aligned} \mathcal{K}^I_p, \mathcal{H}^{i-1}_p &= \mathcal{K}^I W^i_1 + b^i_1, \mathcal{H}^{i-1} W^i_2 + b^i_2 \\ \mathcal{M}^i &= concat(\mathcal{K}^I_p, \mathcal{H}^{i-1}_p) W^i_3 + b^i_3 \end{aligned} \tag{13}$$

where $W^i_1, W^i_2 \in \mathbb{R}^{d\times d}$, $W^i_3 \in \mathbb{R}^{2d\times d}$ and $concat(\cdot)$ is concatenation operation in the embedding dimension.

Encoder-Decoder We use encoder-decoder structure as Language Generation Model based on the Pre-Norm Transformer [45]. Specifically, the encoder is formulated as:

$$\mathcal{S}^I = Encoder(\mathcal{V}^{\mathcal{I}}) \tag{14}$$

where $\mathcal{V}^{\mathcal{I}} \in \mathbb{R}^{m\times d}$ is the visual features extracted by the Visual Extractor (e.g., ResNet [13], ViT [11]), $\mathcal{S}^I \in \mathbb{R}^{m\times d}$, with m patches and embedding dimension d. The decoding process of predicting the current word is:

$$H^i_t = \begin{cases} Decoder(Y_{<t}, \mathcal{S}^I), & \text{if } i = 1 \\ Decoder(\mathcal{M}^i_{<t}, \mathcal{S}^I), & \text{if } i \geqslant 2 \end{cases} \tag{15}$$

where $Y_{<t}$ is the sequence feature of the previously generated sentence before time step t, and H^i_t is the hidden state output by i-th decoder to predict the current word.

3.4. Training Strategy

Global Information Self-Distillation The utilization of ADM to attenuate background information may result in the issue of feature erosion where a small number of critical areas are also weakened. To prevent potential feature erosion, we employ a pre-trained model with a structure identical to EKAGen as a teacher network, taking the entire image as input to distill knowledge into the EKAGen model. In addition, our strategy differs from vanilla knowledge distillation, where the parameters are frozen. Instead, we transfer

the student's weights to the teacher when the student outperforms the teacher network in a new epoch. We use KL divergence as distillation loss to align the probability distributions of the teacher model and EKAGen:

$$\mathcal{L}_{KL} = \frac{1}{N}\sum_{c=1}^{N} KL[p_t(c, I)||p_s(c, I)] \tag{16}$$

where $p_t(c, I)$, $p_s(c, I)$ is the probability distribution of the teacher model and EKAGen, respectively, for word index c and image I. N is the dimensionality of the word space.

Training Loss The language generation process is optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{lg} = -\sum_{t=1}^{n} logp(y_t^*|y_{<t}^*, \mathcal{K}^{\mathcal{I}}, I) \tag{17}$$

where $y_{<t}^*$ denotes the ground truth of the report sequence. $\mathcal{K}^{\mathcal{I}}$ is IEK generated by Equation 6, and I is input image. The final training objective $\mathcal{L}$ is the combination of $\mathcal{L}_{lg}$ and $\mathcal{L}_{KL}$, with $\mathcal{L}_{KL}$ scaled by factor δ:

$$\mathcal{L} = \mathcal{L}_{lg} + \delta\mathcal{L}_{KL} \tag{18}$$

4. Experiments

4.1. Datasets, Metrics and Settings

We evaluate our model on two widely-used benchmarks for report generation: IU X-Ray [8] and MIMIC-CXR [20].

IU X-Ray IU X-Ray dataset, developed by Indiana University, is a widely-used dataset containing 7,470 X-ray images and 3,955 corresponding reports. We follow the established training-validation-testing splits of previous research [33, 47] with a distribution ratio of $70\% : 10\% : 20\%$.

MIMIC-CXR MIMIC-CXR dataset, released by Beth Israel Deaconess Medical Center, is a comprehensive chest X-ray dataset containing 473,057 radiographs and 206,563 corresponding reports. Following previous works [5, 6, 26, 32, 33], we utilize the official split, where the training set consists of 368,960 images, the validation set contains 2,991 images, and the test set comprises 5,159 images.

Metrics We evaluate the quality of the generated reports using widely adopted natural language generation (NLG) metrics including BLEU [37], METEOR [3], and ROUGE-L [30] following the standard evaluation protocol [2]. Following [5, 6, 26, 47, 48], we utilize the CheXpert [19] to label the generated reports and employ Precision, Recall, and F1-Score to evaluate clinical efficacy (CE) metrics. The alignment score is defined as the proportion of cosine similarities greater than 0.5 between the features of report pairs.

Settings Our baseline consists of a pre-trained ResNet-101 and a Pre-Norm transformer with 6 layers, initialized randomly. Our EKAGen offers two versions for the Visual Extractor: ResNet-101 [13] and ViT-B/16 [11]. The transformer has 8 heads and a dimensionality of 256. We utilize swin transformer and ResNet34 as the DiagnosisBot and Anomaly Locator, respectively, to generate diagnostic logits and activation maps. For the generation process of ADM, the threshold value θ is set to 0.6 to filter out the salient regions, and the scaling factor γ is set to 0.4 to suppress the background. The loss $\mathcal{L}_{KL}$ scaling factor δ is set to 0.01. We utilize the AdamW optimizer with learning rates of 1×10^{-5} for the Visual Extractor and 1×10^{-4} for the language generation model. The training batch sizes for MIMIC-CXR and IU X-Ray are set to 32 and 16, respectively. All experiments are run on the RTX 4090 GPU.

[2] https://github.com/tylin/coco-caption

4.2. Quantitative Analysis

Comparison with State-of-the-Art Methods We compare our experimental results with state-of-the-art (SOTA) methods on the IU X-Ray and MIMIC-CXR datasets. The contrastive methods include five categories: Knowledge Based [18, 32, 47, 49], Pre-training [24, 48], Memory Driven [6, 38], Contrastive Based [26, 33], and Image Captioning [7, 56]. As shown in Table 1, EKAGen (RN-101) achieves the SOTA performance across most metrics, such as a 2.6% increase in BLEU-1 and a 0.8% improvement in BLEU-2 on MIMIC-CXR dataset. EKAGen's scores are slightly lower than KiUT [18] and METransformer [47] in a few metrics, possibly due to KiUT's proficiency in synonym representation within the symptom graph and METransformer's word voting strategy favoring phrase alignment.

Analysis on Backbone Table 1 also presents a comparison of the results achieved by EKAGen when utilizing the ResNet-101 [13] and ViT-B/16 [11] visual extractors on the IU X-Ray and MIMIC-CXR datasets. In comparison to ResNet-101, ViT-B/16 exhibited a decrease in performance on both datasets. A potential reason for this observation is that the CNN structure of ResNet-101 is more sensitive to capturing local fine-grained features, which are crucial for the model's understanding of pathological features in the human body and their translation into reports.

Analysis on Clinical Efficacy Metrics The CE metrics are more effective in evaluating the accuracy of pathological description. As shown in Table 2, Our method has shown a significant improvement compared to previous approaches, with Precision, Recall, and F1-Score increasing by 4.6%, 4.8%, and 8.4% respectively. The reason is that EKAGen synthesizes patient conditions, generates expert knowledge at the instance-level, and focuses on pivotal regions to produce reports with more accurate disease assessments.

Analysis on Embedding Network Table 3 presents the experimental results of the unified embedding network in EKAGen and the separate BERT [9] prior knowledge embedding network on the MIMIC-CXR and IU X-Ray

Type	Model	IU X-Ray						MIMIC-CXR					
		BL-1	BL-2	BL-3	BL-4	MTOR	RG	BL-1	BL-2	BL-3	BL-4	MTOR	RG
Image Captioning	M2transformer [7]	0.463	0.318	0.214	0.155	-	0.335	0.212	0.128	0.083	0.058	-	0.240
	Grounded [56]	0.446	0.301	0.237	0.176	-	0.343	0.271	0.174	0.122	0.094	-	0.257
Contrastive Based	CA [33]	0.492	0.314	0.222	0.169	0.193	0.381	0.350	0.219	0.152	0.109	0.151	0.283
	DCL [26]	-	-	-	0.163	0.193	0.383	-	-	-	0.109	0.150	0.284
Memory Driven	R2GenCMN [6]	0.475	0.309	0.222	0.170	0.191	0.375	0.353	0.218	0.148	0.106	0.142	0.278
	R2GenRL [38]	0.494	0.321	0.235	0.181	0.201	0.384	0.381	0.232	0.155	0.109	0.151	0.287
Pre Training	BLIP [24]	0.471	0.294	0.216	0.157	0.186	0.358	0.351	0.215	0.146	0.107	0.151	0.265
	Clinical-BERT [48]	0.495	0.330	0.231	0.170	-	0.376	0.383	0.230	0.151	0.106	0.144	0.275
Knowledge Based	GSKET [49]	0.496	0.327	0.238	0.178	-	0.381	0.363	0.228	0.156	0.115	-	0.284
	PPKED [32]	0.483	0.315	0.224	0.168	-	0.376	0.360	0.224	0.149	0.106	0.149	0.284
	KiUT [18]	0.525	0.360	0.251	0.185	**0.242**	**0.409**	0.393	0.243	0.159	0.113	**0.160**	0.285
	METransformer [47]	0.483	0.322	0.228	0.172	0.192	0.380	0.386	0.250	0.169	**0.124**	0.152	**0.291**
Ours	EKAGen (ViT-B/16)	0.517	0.351	0.258	0.191	0.211	**0.409**	0.415	0.254	0.166	0.117	0.154	0.285
	EKAGen (RN-101)	**0.526**	**0.361**	**0.267**	**0.203**	0.214	0.404	**0.419**	**0.258**	**0.170**	0.119	0.157	0.287

Table 1. Comparing the performance of our EKAGen with other state-of-the-art methods on IU X-Ray and MIMIC-CXR datasets. The comparison scores are cited from the primary publication and paper [46], with the highest performing results highlighted in bold. The abbreviations BL, MTOR, and RG correspond to BLEU, METEOR, and ROUGE, respectively.

MODEL	MIMIC-CXR		
	Precision	Recall	F1-Score
R2GenCMN [6]	0.334	0.275	0.278
GSKET [49]	0.458	0.348	0.371
Clinical-BERT [48]	0.397	0.435	0.415
KiUT [18]	0.371	0.318	0.321
DCL [26]	0.471	0.352	0.373
METransformer [47]	0.364	0.309	0.311
EKAGen (RN-101)	**0.517**	**0.483**	**0.499**

Table 2. The comparison of the clinical efficacy metrics on MIMIC-CXR dataset, with the highest scores highlighted in bold.

Dataset	Embedding	BL-1	BL-2	MTOR	RG
IU X-Ray	BERT	0.507	0.348	0.211	0.399
	Uniform	**0.526**	**0.361**	**0.214**	**0.404**
MIMIC-CXR	BERT	0.409	0.251	0.153	0.276
	Uniform	**0.419**	**0.258**	**0.157**	**0.287**

Table 3. Comparing the performance of a unified prior knowledge encoding network with separate BERT encoding in report generation on the IU X-Ray and MIMIC-CXR datasets.

datasets. Compared with unified embedding, BERT embedding dramatically decreased on both datasets, with metrics such as BLEU-1 and BLEU-2 scores dropping by 1.9% and 1.3% respectively on the IU X-Ray dataset. The experimental observation validates that the unified encoding network of EKAGen can effectively preserve the textual features of prior knowledge during the embedding process, thus averting the problem of feature degradation during decoding.

4.3. Ablation Study

Effect of IEK As shown in Table 4, compared models (a,b,c) with the BASE model reveals that incorporating diagnostic knowledge to specific cases can enhance performance. Compared with the BASE model, (a) achieves a BLEU-1 improvement of 1.2%, while (b) achieved a BLEU-1 improvement of 3.1%. This indicates that (a), which utilizes expert knowledge and performs operations in the embedding space as the knowledge support, can enhance generation but has limited capability. (b) by searching for the nearest features in the original embedding space to form the knowledge support, effectively corrects feature disruption caused by operations. Compared (c) with (b), by accurately retrieving knowledge features in RKS based on patient diagnosis, brings further gains, e.g., 0.494 $\rightarrow$ 0.501 and 0.156 $\rightarrow$ 0.170 in BLEU-1 and BLEU-4 scores.

Effect of ADM During patient examinations, clinical experts typically rely on observations of critical regions as a reference for report writing. ADM simulates this process by leveraging weak supervision to generate activation maps that enhance attention on pivotal regions. In Table 4, models (d) and (c) demonstrate significant performance improvements achieved by our ADM, with BLEU-4 and ROUGE-L scores increased by 2.8% and 1.1% respectively. Figure 4 (a) shows the BLEU-1 scores for different background scaling factors, with the highest score achieved when γ is set to 0.4. These experimental results validate the effectiveness of ADM. By focusing more attention on pivotal regions, our approach can generate smoother descriptions that demonstrate a higher semantic similarity to expert reports.

Effect of GID Models (e) and (d) in Table 4 demonstrate

DATA	SETTING	IEK			ADM	GID	NLG METRICS					
		PKS	RKS	EKN			BL-1	BL-2	BL-3	BL-4	MTOR	RG
IU X-Ray	BASE						0.463	0.287	0.200	0.149	0.178	0.346
	(a)	✓					0.475	0.307	0.209	0.148	0.199	0.365
	(b)	-	✓				0.494	0.319	0.217	0.156	0.199	0.379
	(c)	-	✓	✓			0.501	0.328	0.230	0.170	0.206	0.386
	(d)	-	✓	✓	✓		0.509	0.349	0.259	0.198	0.212	0.397
	(e)	-	✓	✓	✓	✓	**0.526**	**0.361**	**0.267**	**0.203**	**0.214**	**0.404**

Table 4. Quantitative analysis of EKAGen on the IU X-Ray dataset. The BASE model comprises of a Feature Extractor and an Encoder-Decoder structure. The abbreviations BL, MTOR, and RG correspond to the metrics BLEU, METEOR, and ROUGE, respectively.

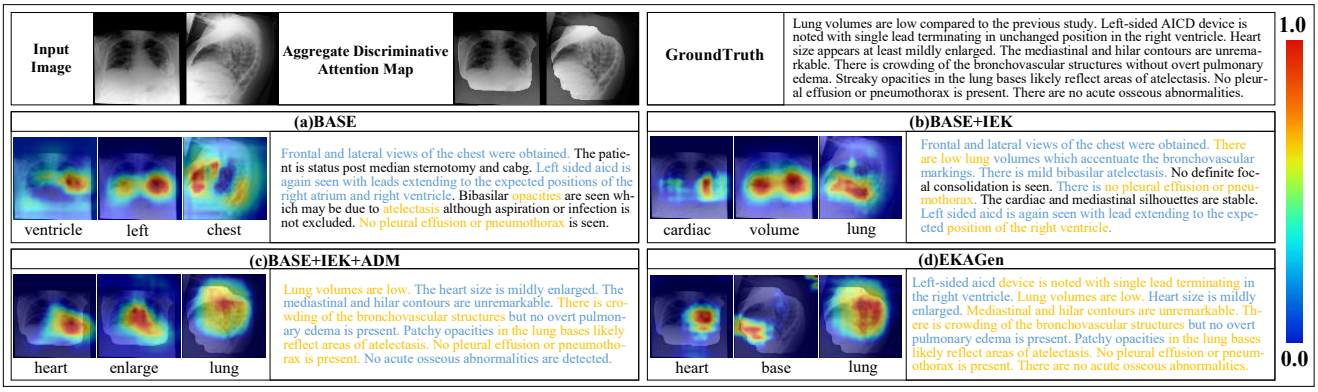

Figure 3. Image-text attention visualizations and captioning results from EKAGen and other models on the MIMIC-CXR dataset. Gold indicates complete alignment with the ground truth, while blue represents semantic alignment.

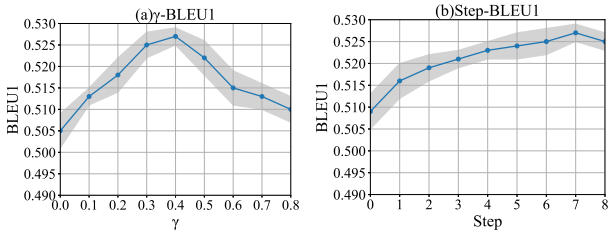

Figure 4. (a) is the impact of scaling factor γ on BLEU-1, while (b) is the variation of BLEU-1 on the GID iterations.

the effectiveness of our Global Information Self-Distillation strategy. By using the model trained on complete images as the teacher network to distill knowledge into our EKAGen, significant improvements in performance are observed, e.g., 0.509 → 0.526 and 0.349 → 0.361 in BLEU-1 and BLEU-2 scores. Figure 4 (b) is the variation of BLEU-1 scores for different iteration steps, with highest score achieved at step 7. This shows that GID, using soft labels, enhances supervision and effectively tackles global information loss, leading to improved accuracy without needing additional labels.

Analysis on alignment scores: To evaluate the similarity of the reports generated by various models with the ground truth on a holistic level, we calculate the alignment scores of different models. As shown in Figure 6 (a), we can observe that BASE, BASE+IEK, BASE+IEK+ADM and EKAGen achieve scores of 0.597, 0.733, 0.752, and 0.785, respectively. This demonstrates EKAGen in implicitly aligning generated features with ground truth features.

4.4. Qualitative analysis

Report Analysis In Figure 3, we present the reports generated by our models. Compared with (a), the model detects more pathological information, such as "low lung volumes", when incorporating instance-level prior knowledge. Furthermore, when using ADM to highlight key regions in (c), the model can detect more abnormal contour such as "heart mildly enlarged". Additionally, by introducing GID to prevent feature loss, the model generates medical device-related terms like "left-sided aicd". Notably, among all settings, the reports generated by EKAGen are the most comprehensive and closest in length to the ground truth.

Attention Visualization Figure 3 visualizes the ADM and image-text attention mapping generated by our models. Compared with (a), after integrating IEK, (b) demonstrates more accurate localization of organ regions (e.g., cardiac, lung). Additionally, (c) and (d) show improved perception of anatomical features such as "heart", through ADM reinforcement of key areas. In (c), the model exhibits increased

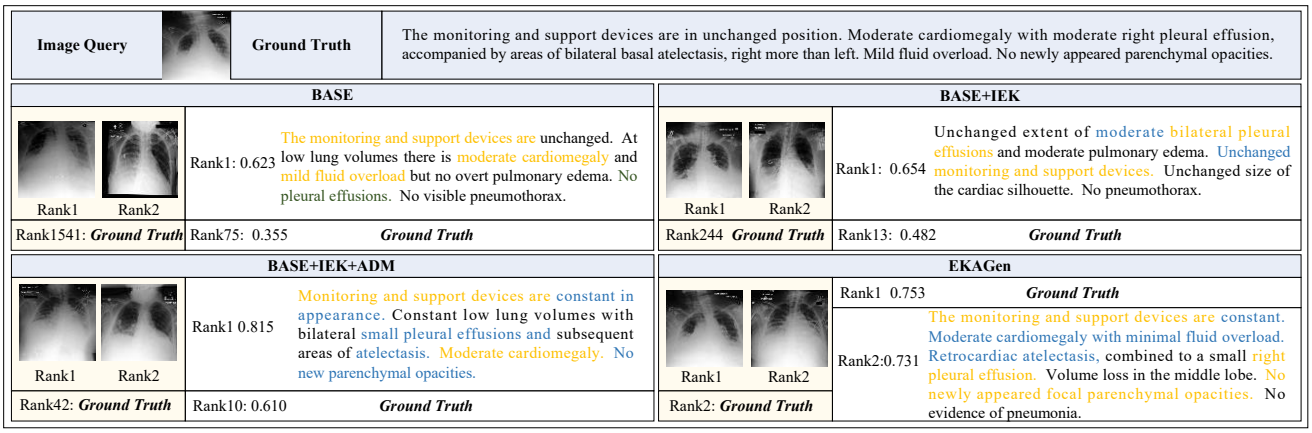

Figure 5. Our models for image retrieval and report retrieval are evaluated using the MIMIC-CXR dataset. In report retrieval, accurate statements are indicated by gold and statements with semantic similarity are highlighted in blue. We utilize ranking and cosine similarity as evaluation metrics for assessing the importance level.

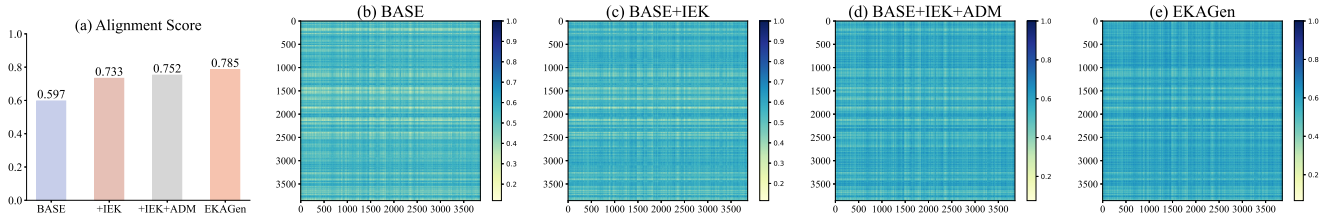

Figure 6. (a) displays the alignment scores of different models, while (b-e) show heatmaps of pairwise cosine similarity among all test samples on the MIMIC-CXR dataset. EKAGen fits better across the entire dataset compared with other models.

sensitivity in capturing anomalies such as "enlarge", while (d) provides more fine-grained supervision through GID's soft label, accurately identifying the "base" region of the lungs and the position of the "heart", enabling the capture of small targets in the ventricle area for a "single lead".

Image and Report Retrieval Figure 5 illustrates the results of image and report retrieval using our EAKGen and other models. Observations reveal that compared with other models, EAKGen exhibits a higher cosine similarity with the ground truth report. By incorporating IEK, ADM, and GID into the BASE model, it exhibits an increased retrieval rank, indicating a growing probability of retrieving the query images and ground truth reports. Additionally, the reports retrieved by EAKGen exhibit a closer resemblance to the ground truth in terms of content. For instance, in the case of EKAGen's rank 2, it accurately identifies and diagnoses "right pleural effusion," providing precise information about both the disease and its location. These findings indicate that our method produces semantic information that better aligns with the ground truth report.

Pairwise Cosine Similarity As depicted in Figure 6, (b) indicates that BASE has a limited number of samples similar to the query sample. Comparing (e) and (b), EKAGen generates diagnostic reports that are closer to those of clinical experts than the BASE. From (c-e) and (b), it is evident that removing IEK, ADM or GID significantly reduces the similarity between generated reports and expert reports. This reinforces our method's ability to narrow the gap between report generation methods and human experts.

5. Conclusion

In this paper, we initially develop comprehensive embedding representations for pulmonary disease and introduce IEK to mitigate the issue of feature degradation. Subsequently, we utilize weak supervision to generate activation maps that highlight crucial regions and create ADM to prioritize discriminative regions. Lastly, we propose the GID strategy to prevent feature erosion and provide soft supervision, distilling global knowledge into our model. Extensive experiments and analyses on the IU X-Ray and MIMIC-CXR datasets validate the efficacy of our EKAGen, which achieves state-of-the-art performance on both datasets.

Acknowledgment

This work was supported by National Natural Science Foundation of China (NSFC) (Grant 92249303), National Key Research and Development Program of China (2023YFF1204900), Natural Science Foundation of Guangdong Province (Grant 2023A1515011907), and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant 23xkjc003).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 2

[3] Satanjeev Banerjee and Alon 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*, pages 65–72, 2005. 5

[4] Shenshen Bu, Taiji Li, and Zhiming Dai. Enhancing medical report generation in multi-slice fusion scenarios. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1030–1037. IEEE, 2023. 2

[5] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 1, 5

[6] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, 2021. 1, 2, 5, 6

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 1, 2, 5, 6

[8] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 1, 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5

[12] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning. In *Proceedings of the IEEE international conference on computer vision*, pages 1222–1231, 2017. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[14] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13450–13459, 2022. 1

[15] Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688, 2022. 2

[16] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022. 1, 2

[17] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 2

[18] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023. 2, 5, 6

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 5

[20] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 5

[21] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979, 2022. 1

[22] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6666–6673, 2019. 2

[23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi.

Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 5, 6

[25] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665, 2022. 1

[26] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun 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*, pages 3334–3343, 2023. 2, 5, 6

[27] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26 (1):253–270, 2023. 1, 2

[28] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[29] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874, 2023. 1, 2

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[31] Fenglin Liu and Wu Xian Yuille Qihang Ge, Shen. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, 2021. Association for Computational Linguistics. 1, 2

[32] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian 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*, pages 13753–13762, 2021. 2, 5, 6

[33] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, 2021. 5, 6

[34] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279, 2021. 1

[35] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. 2

[36] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 2

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[38] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022. 5, 6

[39] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 2

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4

[41] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3

[42] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 2

[43] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[44] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer, 2022. 1

[45] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, 2019. 4

[46] Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, 41(10):2803–2813, 2022. 2, 6

[47] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping 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*, pages 11558–11567, 2023. 1, 2, 5, 6

[48] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2982–2990, 2022. 2, 5, 6

[49] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, page 102510, 2022. 1, 2, 5, 6

[50] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2

[52] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[53] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474, 2021. 2

[54] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12910–12917, 2020. 2

[55] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. 1

[56] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4777–4786, 2020. 5, 6