Dynamic Knowledge Prompt for Chest X-ray Report Generation

Shenshen Bu, Yujie Song, Taiji Li, Zhiming Dai*
School of Computer Science and Engineering, Sun Yat-sen University, China
{bushsh, songyj9, litj5}@mail2.sysu.edu.cn, daizhim@mail.sysu.edu.cn

Abstract

Automatic generation of radiology reports can relieve the burden of radiologist. In the radiology library, the biased dataset and the sparse features of chest X-ray image make it difficult to generate reports. Many approaches strive to integrate prior information to enhance generation, but they fail to dynamically utilize pulmonary lesion knowledge at the instance-level. To alleviate above problem, we propose a novel Dynamic Knowledge Prompt (DKP) framework for chest X-ray report generation. The DKP can dynamically incorporate the pulmonary lesion information at the instance-level to facilitate report generation. Initially, we design a knowledge prompt for each pulmonary lesion using numerous radiology reports. After that, the DKP using an anomaly detector generates the dynamic knowledge prompt by extracting discriminative lesion features in the corresponding chest X-ray image. Finally, the knowledge prompt is encoded and fused with hidden states extracted from decoder, to form multi-modal features that guide visual features to generate reports. Extensive experiments on the public datasets MIMIC-CXR and IU X-Ray show that our approach achieves state-of-the-art performance.

Keywords: Dynamic Knowledge Prompt, Multi-modal Features, Report Generation

1. Introduction

In clinical practice, writing radiology reports can be time-consuming and error-prone when the doctor is in a state of fatigue from working long hours. Automatic radiographic report generation can alleviate the burden of radiologists, especially in critical situations of COVID-19 or other pandemic diseases.

Most existing medical report generation models (Yuan et al., 2019; You et al., 2021) have achieved decent performance in radiology using an encoder-decoder framework. However, medical report generation still suffers from serious problems, **1)Data bias:** Control samples dominate the whole dataset, and the abnormal regions are much smaller than the normal regions in the images of patient samples. Thus, most approaches can learn normal descriptions, but fail to capture anomalies; **2)Visual feature sparsity:** Different from natural images, radiology images lack sufficient discriminative features, which leads to most methods not learning their complex structure and diversity.

To alleviate the above problems, some studies incorporate some prior knowledge to the model. Specifically, some methods (Zhang et al., 2020; Liu et al., 2021b) learned abnormal relationships by applying medical knowledge graphs of certain abnormal conditions. GECL (Hu et al., 2022) integrated additional knowledge and original findings together to extract critical information. GSKET (Yang et al., 2022) fused general and specific knowledge with the visual features of radiology images to facilitate generation. Despite significant advances in these methods, they still have limitations in making full use of the coupling relationship between image and report. For instance, these methods do not utilize the lesion information at the instance level.

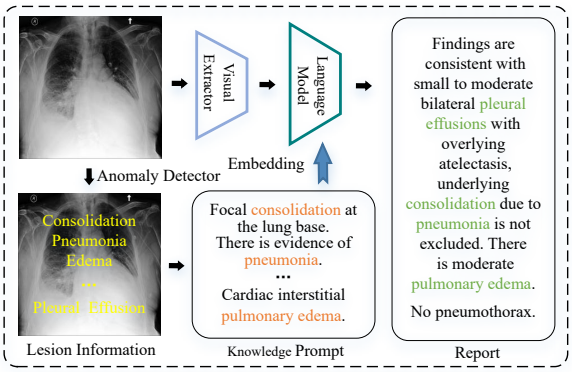

Figure 1: Our proposed DKP generates instance-level knowledge prompt by extracting critical pulmonary lesions information to boost generation.

In radiology image diagnosis, the first step for a radiologist is to focus on abnormal areas and identify lesion locations before creating a comprehensive report. This process can be emulated by using an anomaly detector that offers precise diagnostic information regarding pulmonary lesions.

In this paper, we propose a novel Dynamic Knowledge Prompt (**DKP**) framework for chest X-ray report generation, which can imitate the working patterns of radiologists to generate reports. Inspired by prompt learning (Liu et al., 2021d; Zhou et al., 2022c), by dynamically generating instance-level prompt, DKP is able to provide critical disease category information for report generation. Different from the previous modes of filling in the blanks (Rad-

* Corresponding author

LREC-COLING 2024, pages 5425–5436
20-25 May, 2024.

ford et al., 2021) and adding trainable parameters (Jia et al., 2022; Zhou et al., 2022b) for fine-tuning, DKP contains a new prompt paradigm. Specifically, as shown in Fig. 1, given an image, the DKP first identifies the lesion information at the instance-level, and then generates the corresponding dynamic knowledge prompt, which is combined with hidden states extracted from decoder to boost generation. In order to complete the above process, the DKP introduces three modules, Dynamic Instance Level Explorer (ILE), Prior Knowledge Prompt Fuser (KPF) and Knowledge Distillation Decoder (KDD).

Our contributions are summarized as follows:

- To alleviate data bias, we propose the Dynamic Instance Level Explorer (**ILE**) which dynamically assigns an instance-level prior knowledge prompt for each image to strengthen the DKP's description of abnormal regions.
- To alleviate the sparsity of visual features, we propose the Prior Knowledge Prompt Fuser (**KPF**) which fuses dynamic knowledge prompt with hidden states to create enhanced multi-modal features, thereby enhancing generation.
- To effectively guide report generation, we propose the Knowledge Distillation Decoder (**KDD**), which distills the critical information from multi-modal features to generate reports.

2. Related Works

2.1. Chest X-ray Report Generation

Benefiting from the success of image captioning methods, a significant number of radiology report generation approaches (Zhang et al., 2020; Liu et al., 2021a; Jing et al., 2018; You et al., 2021) were proposed in recent years. R2GenCMN (Chen et al., 2021) introduced a cross-modal memory network that enhances interaction across different modalities. Contrastive learning-based methods (Zhou et al., 2022a; Liu et al., 2021c) were utilized to align feature relationships between images and texts. Furthermore, certain approaches incorporated prior knowledge to address the limited features of radiology images. PPKED (Liu et al., 2021b) introduced a knowledge-enhanced approach by combining visual features with general and specific knowledge to enhance the quality of generated reports. GECL (Hu et al., 2022) used a graph encoder to extract correlations among medical entities and a dependency tree to improve the representation performance of pre-trained text encoders. Although these methods showed improvement in report generation, they did not consider specific prior knowledge related to lesions at the instance-level.

2.2. Prompt Learning

Prompt Learning (Liu et al., 2021d) was proposed and improved by the GPT series (Brown et al., 2020; Radford et al., 2019) in the field of natural language processing, which helps the pretrained model learn about the downstream task. GPT-3 pioneered each downstream task as a mask modeling problem, where the model learns text representations directly within the prompt. Subsequently, a large number of studies (Shin et al., 2020; Jiang et al., 2020) were devoted to developing efficient prompt strategies to extract knowledge from pre-trained large models. Specifically, several approaches (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021e) treated prompts as continuous vectors for specific tasks and optimized them directly during fine-tuning, namely prompt tuning. VPT (Jia et al., 2022) incorporated learnable tokens into the input of the Transformer block, effectively introducing learnable pixels into the input space. In contrast to previous methods such as filling in the blanks (Radford et al., 2021) and adding trainable parameters (Zhou et al., 2022b) for fine-tuning, our DKP introduces a novel prompt paradigm. It dynamically generates instance-level diagnostic knowledge prompts for each chest X-ray image, thereby enhancing radiology report generation.

3. The Proposed Approach

An overview of the proposed model is demonstrated in Fig. 2. As in a previous study (Cho et al., 2021), we unify the report generation task into a text generation problem. Given a radiology image I, the source sequence $X^I = \{x_1, x_2, ..., x_s, ..., x_S\}$, where $x_s \in \mathbb{R}^d$ is extracted from chest X-ray image I by ResNet-101 (He et al., 2016) (parameterized by θ^V). Knowledge prompt sequence $K^I = \{k_1, k_2, ..., k_l, ..., k_L\}$, where $k_l \in \mathbb{W}$ is the knowledge prompt tokens generated by ILE, L is the length of the prompt, and $\mathbb{W}$ is the word space. We aim to maximize the agreement between output of the model and target sequence $Y^I = \{y_1, y_2, ..., y_m, ..., y_M\}$, where $y_m \in \mathbb{W}$ is target tokens, M is target length. We use an encoder-decoder language model (parameterized by θ^{L}) as the main generative model. ILE is used to generate dynamic knowledge prompt during training the generative model, so the weights of ILE are pre-trained and need to be frozen when training the generative model. The Transformer encoder (parameterized by θ^E) is used to encode the source sequence X^I.The KPF (parameterized by θ^F) encodes knowledge prompt and then fuse it with the hidden states output by KDD to form multi-modal features. The KDD (parameterized by θ^{D}) learns the cross-modal representations and utilizes them to generate the radiology report. Optimization ob-

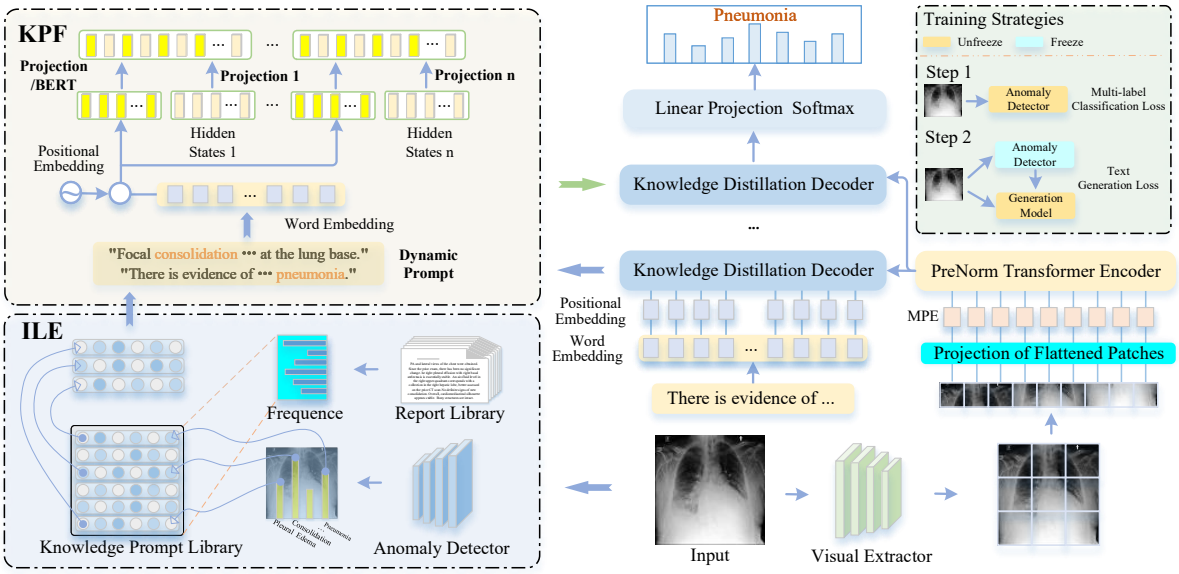

Figure 2: Overview of our DKP architecture. DKP consists of an ILE component for dynamically generating instance-level knowledge prompt, a KPF for generating knowledge-promt-based multi-modal features, a Visual Extractor, a PreNorm Encoder, and a KDD for distilling multi-modal features to generate reports.

jective is to minimize the Cross-entropy (CE) loss:

$$\begin{aligned}\mathcal{L}_{gen}(X^I, K^I, Y^I; \theta^V, \theta^F, \theta^L) \\ = CE(f_{\theta^L}(X^I, X^{I\to H}), Y^I) \\ = -\sum_{i=1}^{M} y_i log(f_{\theta^L(X^I, X^{I\to H})_i})\end{aligned} \tag{1}$$

where f_θ denotes a function parameterized by θ, and $X^I = f_{\theta^V}(I)$. The parameters θ^L of the language model can be divided into two parts: θ^E and θ^D. $X^{I\to H}$ is the fused multi-modal features, that is $X^{I\to H} = f_{\theta^F}(f_{\theta^D}(f_{\theta^E}(X^I)), K^I)$.

3.1. Principles for Prompt Library

Our approach incorporates two components into the knowledge prompts for each disease, forming the knowledge prompt library. One component utilizes templates such as **"the evidence of ..."** which aims to enhance natural language generation metrics. The other component includes disease-specific information like **"pneumonia"** which aims to improve clinical efficacy metrics. By incorporating these components, we address both the issue of **smooth readability** and the problem of accurate **medical diagnosis**. Furthermore, it is worth noting that individuals may simultaneously suffer from multiple diseases. Thus, the knowledge prompt for a particular patient may consist of a **combination of prompts** related to different diseases.

3.2. Dynamic Instance Level Explorer

In our approach, ILE is utilized to dynamically generate knowledge prompts for each image. Specifically, we define $T = \{t_1, t_2, ..., t_N\}$ as the knowledge prompt library, where N is the total number of lesions. To dynamically align prior knowledge with each instance, we employ an anomaly detector to identify pulmonary lesions:

$$p = \sigma(AnomalyDetector(I)) \tag{2}$$

where $AnomalyDetector(\cdot)$ is Swin Transformer (Liu et al., 2021f) and σ is sigmoid function. The binary cross-entropy loss is used for optimization:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[l_i log(p_i) + (1-l_i) log(1-p_i)] \tag{3}$$

where $l_i \in \{0, 1\}$ and p_i are the ground-truth and predicted values for the i-th disease, respectively. After applying a threshold, we convert p_i into one-hot encoding to obtain $O^I = \{o_1, o_2, ..., o_N\} \in \mathbb{R}^{N\times N}$, where $o_n \in \mathbb{R}^N$ represents the one-hot vector for the n-th lesion. The prompt scheme $Pro(o_n, T)$ for the n-th lesion is defined as:

$$Pro(o_n, T) = o_n \odot \{t_1, t_2, ..., t_N\} \tag{4}$$

where $\odot$ is the Hadamard Product. We combine the N prompts into a list and remove empty items:

$$\begin{aligned}Pro^I = [Pro(o_1, T), ..., Pro(o_N, T)] \\ K^I = f_{joint}(f_{sample}(Pro^I))\end{aligned} \tag{5}$$

where $f_{sample}(\cdot)$ is used to remove empty items. Finally, all prompts are concatenated using $f_{joint}(\cdot)$ and tokenized to obtain $K^I = \{k_1, k_2, ..., k_L\}$, where L represents the number of words.

3.3. Prior Knowledge Prompt Fuser

The KPF is applied to encode the knowledge prompt and merge it with the hidden states generated by the KDD, resulting in multi-modal features. Initially, each token of K^I is embedded using a trainable embedding matrix:

$$P_e = \{k_1^{idx}, k_2^{idx}, ..., k_L^{idx}\}W^E \tag{6}$$

where $P_e \in \mathbb{R}^{L\times d}$, $W^E \in \mathbb{R}^{m\times d}$, L is the number of tokens, d is the embedding dimension, and m is the size of word space. The parameters of W^E and the weights of the KDD word embedding layer are shared, which can maintain the consistency of prompt and target report information. Then, the learnable position embedding matrix $P_{pos} \in \mathbb{R}^{L\times d}$ is utilized to positional embedding:

$$P = P_e + P_{pos} \tag{7}$$

where $P = \{p_1, p_2, ..., p_L\} \in \mathbb{R}^{L\times d}$. Finally, both the hidden states $hidden^I$ and knowledge prompt features are projected and fused together:

$$\begin{aligned} K_{hid} &= f_h(\{h_1, h_2, ..., h_M\}) \\ K_{pro} &= f_p(\{p_1, p_2, ..., p_L\}) \\ F_{mix} &= fusion(K_{hid}, K_{pro}) \\ K^{mix} &= F_{mix}W_{reduce} \end{aligned} \tag{8}$$

where $f_h(\cdot)$ and $f_p(\cdot)$ are mapping functions for hidden states and prompt features, respectively. We explored different attention mechanisms, including channel attention similar to SENet (Hu et al., 2018) and CBAM (Woo et al., 2018), and logic gate structures similar to LSTM. However, we find that linear projection and BERT (Devlin et al., 2018) performed better. Thus, we propose two versions for our approach i.e. linear projection and BERT. $W_{reduce} \in \mathbb{R}^{2d\times d}$ is used to reduce the dimension of the fused features from $2d$ to d.

3.4. Visual Feature Extraction

Given a radiology image I, the features are extracted by ResNet-101, and the output of the last convolutional layer is reduced by the projection matrix W^I to generate a series of non-overlapping patches. Then the patches are expanded and subsequently connected as input to the position encoding layer. The process is formulated as:

$$X^I = Flatten(ResNet(I)W^I) \tag{9}$$

where $ResNet(I) \in \mathbb{R}^{h\times w\times 2048}$, $W^I \in \mathbb{R}^{2048\times d}$ is used to reduce the dimension from 2048 to d, and $X^I \in \mathbb{R}^{S\times d}$, where $S = h \times w$, $d = 256$. To retain the positional information of image patches, we incorporate position embeddings. Given that the contour of an X-ray image is typically fixed, similar to (Vaswani et al., 2017), we introduce **Matrix Absolute Position Embedding(MPE)**:

$$\begin{aligned} PE_{2i} &= \begin{cases} sin(\mathscr{X}_{pos}/10^{8i/d}) & i < d/2 \\ sin(\mathscr{Y}_{pos}/10^{8i/d}) & i \geqslant d/2 \end{cases} \\ PE_{2i+1} &= \begin{cases} cos(\mathscr{X}_{pos}/10^{8i/d}) & i < d/2 \\ cos(\mathscr{Y}_{pos}/10^{8i/d}) & i \geqslant d/2 \end{cases} \end{aligned} \tag{10}$$

where $\mathscr{X}_{pos}$ is the index of the horizontal patch, and $\mathscr{Y}_{pos}$ is the index of the vertical patch. d is the positional embedding dimension. We add the absolute position embedding $PE \in \mathbb{R}^{S\times d}$ to X^I, resulting in $\widetilde{X}^I = \{\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_S\} \in \mathbb{R}^{S\times d}$, and use it as input to subsequent PreNorm Encoder.

Algorithm 1 DKP

Input: Image I, knowledge prompt library T, number layers of encoder and decoder N
Output: The Report G^I

$\widetilde{X}^I = Flatten(ResNet(I)W^I) + PE$
$V^I = PreNormEnc(\widetilde{X}^I)$ // Visual features
$O^I = OH(Sig(AnomalyDet(I)))$ // Lesions
$K^I = f_{joint}(f_{sample}(O^I \odot T))$ // Prompts
$P_e = K^I W^E$ // Embed knowledge prompt
$g_0 = < bos >, t = 0$ // Initialize
while $g_t \neq < eos >$ **do**
  $t = t + 1$
  $G_{<t} = (g_0, ..., g_{t-1})$
  $H_1^I = Dec(V^I, G_{<t}W^E + P_{pos})$
  for $i = 2$ to N **do**
    $K_i^{mix} = [f_h(H_{i-1}^I), f_p(P_e + P_{pos})]W_r$
    $H_i^I = Dec(V^I, K_i^{mix})$
  end for
  $g_t = Softmax(Linear(H_N^I))$
end while
return $G^I = \{g_0, ..., g_t\}$

3.5. PreNorm Encoder

Our pre-experiments show that it is difficult for vanilla Transformer to converge to optimal results on our task. Following (Wang et al., 2019), DKP encodes the image features $\widetilde{X}^I$ through Pre-Norm encoder, can be formalized as:

$$\begin{aligned} \hat{X}^I &= LN(\widetilde{X}^I) \\ X^{I'} &= LN(MHA(\hat{X}^I, \hat{X}^I) + \hat{X}^I) \\ V^I &= FFN(X^{I'}) + X^{I'} \end{aligned} \tag{11}$$

where $LN(\cdot)$ is the Layer Normalization, and $V^I = \{v_1, v_2, ..., v_S\} \in \mathbb{R}^{S\times d}$ is the encoded visual features, which will be guided by the knowledge features to generate a report using KDD.

3.6. Knowledge Distillation Decoder

To further distill the knowledge information from K^{mix}, we propose the KDD component. The se-

quence input to the first layer of KDD, data preprocessing consistent with KPF, is first encoded by the learnable parameter matrix W^E and added to the learnable position embedding matrix $P_{pos} \in R^{L\times d}$. After the second layer, instead of using hidden states decoding, KDD measures the contribution of visual signals $V^I \in \mathbb{R}^{S\times d}$ and knowledge hidden states $hidden^I$ for word prediction. For the prediction of a word w_t at t time steps, the above procedure can be formulated as:

$$D_n(t) = \begin{cases} Dec(V^I, G_{<t}), & \text{if } n = 1 \\ Dec(V^I, K^{mix}_{<t}), & \text{if } n \geqslant 2 \end{cases} \quad (12)$$

where V^I is output of the last layer PreNorm encoder, and n is the current layer id of KDD. $G_{<t} = (g_0, g_1, ..., g_{t-1}) \in \mathbb{R}^{t\times d}$ and $K^{mix}_{<t}$ are the sequence feature and the multi-modal knowledge hidden states of the partial generated report from t time steps ago, respectively. The basic decoding operation $Dec(V^I, G_{<t})$, which is consistent with our PreNorm encoder and uses the PreNorm decoder, can be formulated as follows:

$$\begin{aligned} \hat{G}_{<t} &= LN(MHA(\widetilde{G}_{<t}, \widetilde{G}_{<t}) + \widetilde{G}_{<t}) \\ G'_{<t} &= LN(MHA(\hat{G}_{<t}, V^I) + \hat{G}_{<t}) \\ hidden^I &= FFN(G'_{<t}) + G'_{<t} \end{aligned} \quad (13)$$

where $\widetilde{G}_{<t} = LN(G_{<t})$, and $hidden^I \in \mathbb{R}^{t\times d}$ is the hidden states. Algorithm 1 summarizes the whole generation process of our DKP.

Two-Stage Training Strategy(OST): We introduce two training objectives for DKP models: Eq. 1 for the language generation model and Eq. 3 for ILE. The OST train the ILE exclusively in the first stage and freeze it during the training of language model in the second stage.

4. Experiments

4.1. Datasets

IU X-Ray (Demner-Fushman et al., 2016) contains 7,470 X-ray images and 3,955 radiology reports, which is a baseline dataset widely used to evaluate radiology report generation methods. Consistent with past studies (Qin and Song, 2022), we exclude the sample without findings in the dataset and get 6,471 images and 3,336 reports. For dataset splitting, we use the same splits as (Chen et al., 2021), where training/validation/testing is $70\%/10\%/20\%$ respectively. Finally, the report is preprocessed by tokenizing and removing non-alpha tokens.

MIMIC-CXR (Johnson et al., 2019) is a large chest X-ray dataset containing 377,110 chest X-ray images and 227,835 radiology reports from 64,588 patients. Following (Liu et al., 2021b), we take the official splits to evaluate our approach. Therefore, the training set contains 368,960 samples, the validation set contains 2,991 samples, and the test set contains 5,159 samples. We convert all tokens of the radiology reports to lower-case, remove tokens that occur less than 10 times in the training set and special character tokens that are not commonly used, and finally get 4253 words.

4.2. Experimental Setup

For both datasets, patch features with dimension $19 \times 19 \times 2048$ is extracted by ResNet-101 (He et al., 2016) pre-trained on ImageNet. The extracted features are further projected to 384×256 for the input of subsequent modules. The PreNorm encoder and decoder have 8 heads and 256 hidden dimensions. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) with learning rates of 1×10^{-5} and 1×10^{-4} for the Visual Extractor and language generation model, respectively. In the ILE module, the Tiny Swin Transformer (Liu et al., 2021f) is used as anomaly detector, trained by an AdamW optimizer with a learning rate of 1×10^{-5} and batch size of 64. For the KPF module, the trainable word embedding matrix and position embedding matrix are parameter shared with KDD, and 128 is set as the maximum length of knowledge prompt. The mapping function $f_p(\cdot)$ for knowledge prompts is implemented through two approaches: utilizing BERT and linear mapping, corresponding to DKP-BERT and DKP-Projection models, respectively. The batch size on MIMIC-CXR and IU X-Ray datasets is 64 and 16, respectively, with resolution of 300×300, and trained over 50 epochs. All experiments are conducted on a Tesla A100GPU with 40GB of VRAM for training and testing.

4.3. Comparison Methods

The comparative methods include METransformer (Wang et al., 2023) with multi-expert tokens, DCL(Li et al., 2023) and GSKET (Yang et al., 2022) encoding prior knowledge through graph neural networks, Clinical-BERT (Yan and Pei, 2022) and BLIP (Li et al., 2022) based on pre-trained language models, R2GenCMN (Chen et al., 2021) and R2Gen (Chen et al., 2020) utilizing memory matrices, CMCL (Liu et al., 2021a) employing multimodal curriculum learning, and PPKED (Liu et al., 2021b) leveraging posterior and prior knowledge. The BLIP scores are from fine-tuning on two datasets, while the scores of others are cited from their original papers.

4.4. Evaluation Metrics

We evaluate performance using natural language generation (NLG) metrics[1], clinical efficacy (CE)

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

DATA	MODEL	Pub.	NLG METRICS						CE METRICS		
			BL-1	BL-2	BL-3	BL-4	MTOR	RG	P	R	F1
IU X-Ray	R2Gen	EMNLP20	0.470	0.304	0.219	0.165	0.187	0.371	-	-	-
	PPKED	CVPR21	0.483	0.315	0.224	0.168	-	0.376	-	-	-
	CMCL	ACL21	0.473	0.305	0.217	0.162	0.186	0.378	-	-	-
	R2GenCMN	ACL21	0.475	0.309	0.222	0.170	0.191	0.375	-	-	-
	BLIP	ICML22	0.471	0.294	0.216	0.157	0.186	0.358	-	-	-
	GSKET	MIA22	0.496	0.327	0.238	0.178	-	0.381	-	-	-
	Clinical-BERT	AAAI22	0.495	0.330	0.231	0.170	-	0.376	-	-	-
	DCL	CVPR23	-	-	-	0.163	0.193	0.383	-	-	-
	METransformer	CVPR23	0.483	0.322	0.228	0.172	0.192	0.380	-	-	-
	DKP-BERT	OURS	0.503	0.339	0.241	0.178	0.195	0.392	-	-	-
	DKP-Projection	OURS	**0.507**	**0.344**	**0.245**	**0.181**	**0.214**	**0.398**	-	-	-
MIMIC -CXR	R2Gen	EMNLP20	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	PPKED	CVPR21	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	CMCL	ACL21	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
	R2GenCMN	ACL21	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	BLIP	ICML22	0.351	0.215	0.146	0.107	0.151	0.265	-	-	-
	GSKET	MIA22	0.363	0.228	0.156	0.115	-	0.284	0.458	0.348	0.371
	Clinical-BERT	AAAI22	0.383	0.230	0.151	0.106	0.144	0.275	0.397	0.435	0.415
	DCL	CVPR23	-	-	-	0.109	0.150	0.284	0.471	0.352	0.373
	METransformer	CVPR23	0.386	0.250	0.169	**0.124**	0.152	**0.291**	0.364	0.309	0.311
	DKP-BERT	OURS	0.412	0.254	0.166	0.115	0.156	0.277	0.487	0.452	0.469
	DKP-Projection	OURS	**0.418**	**0.260**	**0.172**	0.120	**0.159**	0.287	**0.496**	**0.461**	**0.478**

Table 1: Our proposed DKP is compared with previous state-of-the-art methods on IU X-Ray and MIMIC-CXR datasets. The best scores are in bold face. BL, MTOR and RG refer to BLEU, METEOR and ROUGE, respectively.

metrics[2] and AUROC. The NLG metrics include BLEU (Papineni et al., 2002), METEOR, and ROUGE-L (Lin, 2004) and CE metrics include Precision, Recall, and F1-Score. The CheXpert[3] is an annotation tool used by the MIMIC official to generate classification ground-truth. Stanford's work (Irvin et al., 2019) showed that the tool's accuracy can reach the level of human experts. We use CheXpert to label the reports generated by our method and compare them with the ground-truth at 14 image labels in MIMIC-CXR dataset. In addition, AUROC is used to evaluate the prediction of the generated reports on keywords related to "*tissue*", "*location*", "*extent*", and "*surgery*". To be more objective, we do not train the classifier separately and define the presence of keywords in both generated report and ground-truth as a correct prediction. The selected keywords do not include disease-related or ambiguous terms that could lead to misdiagnosis. For generation we use the Micro-average CE metric, and for keyword prediction we use two CE metrics, Micro-average and Macro-average. In addition, the alignment quality is assessed using the alignment score, which is defined as the fraction of radiograph-report pairs with feature cosine similarity greater than 0.5.

[2]Note that CE metrics only apply to MIMIC-CXR because the labeling schema of CheXpert is designed for MIMIC-CXR, which is different from that of IU X-Ray.

[3]https://github.com/ncbi-nlp/NegBio

5. Results and Analyses

5.1. Comparisons with Previous Studies

We compare our approach with a wide range of state-of-the-art radiology report generation and image captioning models. As shown in Table 1, our **DKP-Projection** and **DKP-BERT** outperform previous methods nearly in all metrics. Notably, DKP-Projection achieves the highest score. Therefore, we focus solely on DKP-Projection (referred to as DKP hereafter) to describe the experimental results. Specifically, for the MIMIC-CXR dataset, compared with the suboptimal method METransformer (Wang et al., 2023), our **DKP** improves by 3.2%, 1.0% respectively in BLEU1 and BLEU2 metrics. Similarly, the corresponding metrics are improved by 1.1% and 1.4% respectively in the IU X-Ray dataset. Notably, our **DKP** has a significant improvement compared with the suboptimal results in terms of CE metrics. The Precision, Recall and F1-Score are improved by 2.5%, 2.6% and 6.3% respectively. The superior performance of the DKP algorithm lies in its ability to dynamically generate and provide

DATA	SETTING	MODEL				NLG METRICS						CE METRICS		
		MPE	KPF	ILE	OST	BL-1	BL-2	BL-3	BL-4	MTOR	RG	P	R	F1
IU X-Ray	BASE				-	0.463	0.287	0.200	0.149	0.178	0.346	-	-	-
	wo/ILE	✓	✓		-	0.489	0.306	0.214	0.156	0.186	0.389	-	-	-
	wo/KPF	✓		✓	-	0.483	0.312	0.224	0.166	0.194	0.390	-	-	-
	wo/MPE		✓	✓	-	0.492	0.319	0.220	0.159	0.207	0.375	-	-	-
	DKP	✓	✓	✓	-	**0.507**	**0.344**	**0.245**	**0.181**	**0.214**	**0.398**	-	-	-
MIMIC -CXR	BASE					0.378	0.223	0.145	0.101	0.140	0.262	0.429	0.348	0.385
	wo/ILE	✓	✓		-	0.395	0.245	0.160	0.112	0.151	0.276	0.487	0.441	0.463
	wo/KPF	✓		✓	✓	0.392	0.238	0.157	0.111	0.150	0.271	0.493	0.420	0.454
	wo/MPE		✓	✓	✓	0.404	0.242	0.155	0.107	0.148	0.272	0.495	0.449	0.472
	wo/OST	✓	✓	✓		0.398	0.245	0.161	0.112	0.158	0.274	0.489	0.443	0.470
	DKP	✓	✓	✓	✓	**0.418**	**0.260**	**0.172**	**0.120**	**0.159**	**0.287**	**0.496**	**0.461**	**0.478**

Table 2: Ablation experiments of the proposed approach on NLG and CE metrics. The best scores are in bold face and "wo" is defined as "without".

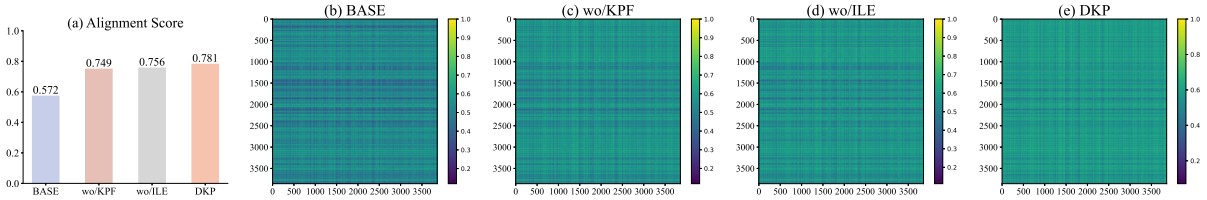

Figure 3: (a) shows the alignment scores of various models, while (b-e) respectively present the heatmaps of pairwise cosine similarity for **BASE**, **wo/KPF**, **wo/ILE**, and **DKP** on the MIMIC-CXR dataset.

instance-level knowledge prompts to the language generation model, using them as key diagnostic information to produce radiology reports with higher accuracy in clinical disease diagnosis.

5.2. Quantitative Analysis

Effect of Dynamic Instance Level Explorer: Compared **DKP** with **wo/ILE** as shown in Table 2, we can find that the performance of the approach decreases significantly if the ILE is removed, e.g., 0.507 → 0.489 and 0.418 → 0.395 in BLEU1 on IU X-Ray and MIMIC-CXR datasets, respectively. For the CE metrics on the MIMIC-CXR dataset, the Precision decreases from 0.496 to 0.487, and the Recall drops from 0.461 to 0.441 upon the removal of the ILE module. Additionally, we find that the removal of ILE resulted in a decrease in the alignment score, e.g., 0.781 → 0.756 on MIMIC-CXR dataset, as shown in Fig. 3(a). The experimental results above indicate that the ILE module significantly enhances language fluency, improves disease diagnosis accuracy, and ensures consistency between algorithm-generated reports and expert reports. The underlying reason for this observation is that the ILE module effectively utilizes the anomaly detector to extract the most discriminative disease-related features and incorporates them into dynamically generated knowledge prompts. This process helps to minimize the disparity between algorithm and expert reports.

Effect of Prior Knowledge Prompt Fuser: After removing KPF, similar to ILE, we observe a significant drop in performance metrics, e.g., 0.507 → 0.483 and 0.418 → 0.392 in BLEU1 on IU X-Ray and MIMIC-CXR datasets. This underscores the role of the KPF module in introducing prior knowledge prompts through direct integration of expert reports on different diseases, thereby enhancing the fluency of algorithm-generated reports. In addition, after removing KPF, the CE metrics show that the Precision decreases slightly, e.g., 0.496 → 0.493, but the Recall decreases significantly, e.g., 0.461 → 0.420. This is consistent with our intuition that the KPF module can significantly improve the Recall of report generation by introducing high frequency descriptions of various lesions.

Effect of Matrix Absolute Position Embedding: Removing MPE leads to a decrease in both NLG and CE metrics. For instance, from 0.507 to 0.492 and from 0.418 to 0.404 in BLEU1 on the IU X-Ray and MIMIC-CXR datasets, respectively. The experimental results demonstrate that, in application scenarios of chest radiology images where the organization of organs is relatively fixed, the Matrix Absolute Position Embedding proposed in this paper exhibits advantages over trainable parameter-based position embedding.

Effect of Two-Stage Training strategy: The experimental comparison between **DKP** and **wo/OST** in Table 2 indicates that the OST strategy outper-

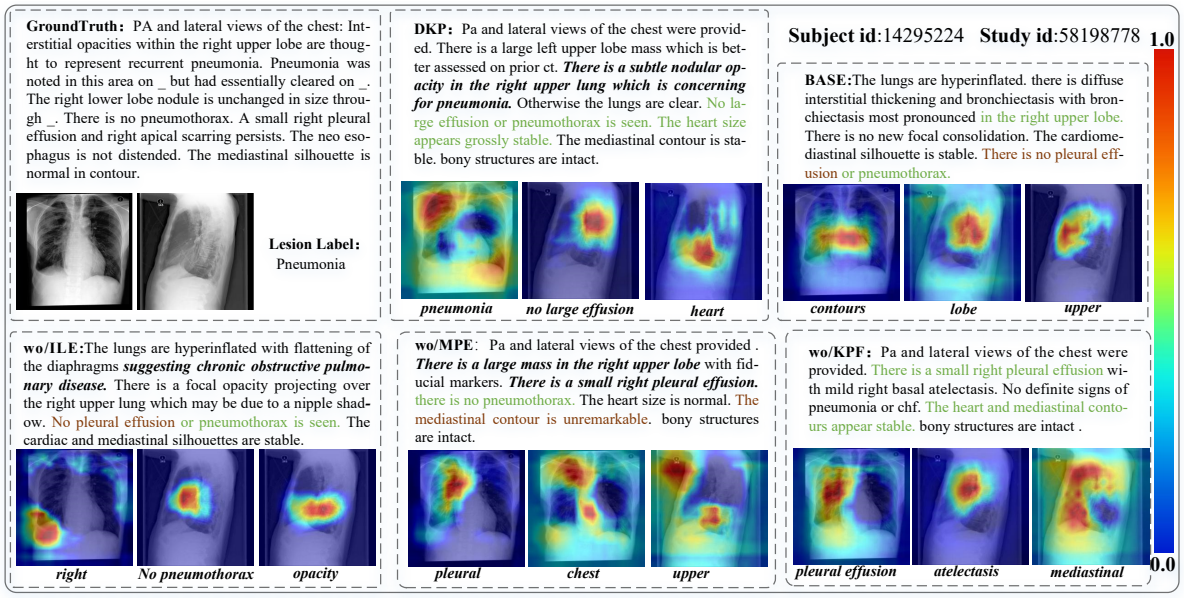

Figure 4: Image-text attention visualizations and case-specific descriptive results from DKP and other baselines. Bold italics and green fonts indicate correct descriptions of the lesion and normal regions, respectively. The red font indicates the error description.

	Desc	MODEL			
		BASE	wo/KPF	wo/ILE	DKP
Tis	tube	0.748	0.671	**0.839**	0.711
	atrium	0.664	0.688	0.697	**0.710**
	gastric	0.560	0.620	**0.749**	0.635
	ventricle	0.821	0.766	**0.853**	0.781
Loc	median	0.699	**0.770**	0.686	0.747
	right	0.602	0.641	0.619	**0.657**
	lateral	0.540	0.556	0.537	**0.562**
	above	0.602	**0.668**	0.552	0.634
Ext	enlarged	0.600	0.603	**0.606**	**0.606**
	opacities	0.526	0.539	0.521	**0.560**
	clear	0.618	0.623	0.619	**0.627**
	moderate	0.549	**0.618**	0.571	0.581
Sur	sternotomy	0.745	**0.800**	0.722	0.785
	devices	0.525	0.729	0.539	**0.788**
	pacemaker	0.611	0.703	**0.754**	0.716
	cabg	0.676	**0.734**	0.704	0.732

Table 3: Comparison AUROC of keywords prediction in MIMIC-CXR. **Tis**, **Loc**, **Ext**, and **Sur** are "*tissue*", "*location*", "*extent*", and "*surgery*".

forms the single-stage training strategy in terms of both NLG and CE metrics. This improvement is attributed to the two-stage training approach, which allows for a clear optimization focus. The first stage targets disease diagnosis classification, while the second stage focuses on report generation, leading to a more stable training process for the algorithm.

Analysis on Keywords Prediction: The AUROC metrics and CE metrics for keyword prediction results on the MIMIC-CXR dataset are shown in Table 3 and Table 4. As we can see, our method

	METRICS	MODEL			
		BASE	wo/KPF	wo/ILE	DKP
Macro	Precision	0.412	0.456	0.419	**0.465**
	Recall	0.358	0.430	0.413	**0.446**
	F1-Score	0.352	0.424	0.386	**0.441**
Micro	Precision	0.454	0.491	0.490	**0.506**
	Recall	0.374	0.407	0.405	**0.435**
	F1-Score	0.410	0.445	0.444	**0.468**

Table 4: The DKP is compared with the prediction results of 16 keywords in MIMIC-CXR.

achieves the best or suboptimal AUROC scores for most keywords, and significantly improves the CE metrics. In addition, by comparing **wo/ILE** of Table 3 with **wo/KPF** and **wo/ILE** of Table 4, KPF effectively promotes the description of *human tissues*, but at the same time, introduces noise, resulting in a decline in CE metrics, which again shows the importance of ILE module. Notably, the **DKP**, which combines ILE and KPF, shows a significant improvement in word scores related to *human tissues* and *surgery*, indicating that our method can effectively facilitate the description of *human tissues* and *surgical treatments*, potentially providing recommendations for therapeutic schemes.

Analysis on Alignment Scores: We assess the similarity between reports generated by different models and the ground truth using alignment score. As illustrated in Fig. 3 (a), the respective scores achieved by **BASE**, **wo/KPF**, **wo/ILE**, and **DKP** are 0.572, 0.749, 0.756, and 0.781. These results demonstrate the ability of DKP to implicitly align generated features with the ground truth.

5.3. Qualitative Analysis

Analysis on Generated Report: In Fig. 4, we present the generated results of DKP along with a comparison to the results of other methods. For the descriptions of radiology image, only **DKP** correctly captures the anomaly information "*pneumonia*" among all methods, and **DKP** describes more comprehensively and most closely to the length of the report written by the radiologist. **wo/ILE** crudely integrates all knowledge prompts, and although it successfully predicts "*pulmonary disease*", it also introduces other noise, which leads to the misdescription of "*no pleural effusion*". This illustrates the effectiveness of our ILE component in successfully sampling the correct knowledge prompt and filtering out noise. Similarly, **wo/KPF** successfully predicts "*small pleural effusion*" compared with **BASE** by adding visual features. However, since the disease category is not explicitly specified, **wo/KPF** does not accurately diagnose "*pneumonia*", demonstrating the effectiveness of KPF module in providing a specific textual knowledge prompt. Replacing Matrix Absolute Position Embedding with trainable position embedding, **wo/MPE** generates an erroneous description of "mediastinal contour is unremarkable". This observation demonstrates the superior effectiveness of MPE in accurately marking the orientation and contour information. **BASE** only manages to identify the location information of the lesion, and similar to **wo/ILE**, it makes an incorrect prediction of "*no pleural effusion*".

Attention Visualization: As shown in Fig. 4, we use Grad CAM (Selvaraju et al., 2017) to visualize image-text attention maps. Our **DKP** accurately focuses on the corresponding location when making predictions on the text, such as the location of "*pneumonia*" or vital organ "*heart*". Similarly, **BASE** and **wo/ILE** also exhibit strong localization capabilities, identifying keywords related to human tissue or contours, such as "contours" and "lobe," orientation-related keywords like "right" and "upper," and symptom-related keywords including "opacity." While **wo/KPF** can generate descriptions of potential diseases, such as "small pleural effusion," it fails to accurately delineate the extent of the disease area "small," further emphasizing the importance of the KPF component in providing clear disease description information to eliminate noise. From **wo/MPE**, it can be observed that after removing MPE, the overall contour, such as localization of "lung", is not particularly distinct. Additionally, introducing noise when localizing phrases like "upper" indicates that MPE possesses better localization capabilities than trainable position embedding.

Pairwise Cosine Similarity Analysis: In Fig. 3 (b-e), we present heatmaps that display the pairwise cosine similarity among all test samples in the MIMIC-CXR dataset. It can be observed from (b) that the **BASE** model exhibits a sparse presence of samples that share similarities with the query sample. Given that radiology reports inherently contain patterns, we would expect to frequently observe high correlations. By comparing (e) and (b), it becomes evident that the **DKP** generates diagnostic reports that closely resemble those created by clinical experts, surpassing the similarity achieved by the **BASE** model. From subplots (c) and (d), it is clear that excluding KPF or ILE results in a substantial reduction in the resemblance between the reports generated by the algorithm and those written by experts. This further strengthens the notion that by simulating the expert diagnostic process, our method is capable of identifying critical disease-related information, dynamically generating knowledge prompts, and then generating a report.

6. Conclusion

In this paper, to alleviate the data bias and visual feature sparse issues, we propose a novel prompt learning paradigm for report generation, which can dynamically generate instance-level knowledge prompts for different cases to boost generation. Extensive experiments and analysis on MIMIC-CXR and IU X-Ray datasets verify the effectiveness of our method. Concretely, DKP effectively improves the quality of radiology report generation by incorporating dynamic knowledge prompt and achieves state-of-the-art performance on both datasets.

Ethics Statement

The providers of the IU X-Ray dataset employ effective techniques to de-identify the text reports, ensuring that the data is anonymized. Consequently, our model does not disclose any details about the patient's identity. Similarly, the MIMIC-CXR dataset does not contain any identifiable information, such as the patient's name, age, or address. Additionally, the reports within these datasets undergo anonymization procedures, ensuring that no patient identity information is revealed. Thus, our model diligently upholds patient confidentiality.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

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

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

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

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

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

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. 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.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2577–2586.

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. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2303.10323*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for 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 4582–4597.

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

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. 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)*. Association for Computational Linguistics.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. 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.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021d. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021e. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021f. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. 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.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

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

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

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–82. Springer.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer.

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

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022c. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.