DDGIP: Radiology Report Generation Through Disease Description Graph and Informed Prompting

Chentao Huang[1], Guangli Li[1(✉)], Xinjiong Zhou[1], Yafeng Ren[2], Hongbin Zhang[1]
[1]East China Jiaotong University, China
[2]Guangdong University of Foreign Studies, China
✉1333@ecjtu.edu.cn

Abstract

Automatic radiology report generation has attracted considerable attention with the rise of computer-aided diagnostic systems. Due to the inherent biases in medical imaging data, generating reports with precise clinical details is challenging yet crucial for accurate diagnosis. To this end, we design a disease description graph that encapsulates comprehensive and pertinent disease information. By aligning visual features with the graph, our model enhances the quality of the generated reports. Furthermore, we introduce a novel informed prompting method which acts as an implicit bag-of-words planning for surface realization, increasing the accuracy of short-gram predictions. Notably, the informed prompting succeeds with a three-layer decoder, reducing the reliance on conventional prompting methods that require extensive model parameters. Extensive experiments on two widely-used datasets, IU-Xray and MIMIC-CXR, demonstrate that our method outperforms previous state-of-the-art models.[1]

1 Introduction

The heavy workload faced by radiologists frequently results in reduced diagnostic accuracy. However, recent breakthroughs in image captioning have paved the way for automated Radiology Report Generation (RRG), providing a promising solution to this pressing issue. A large body of research(Chen et al., 2020; Song et al., 2022; Jin et al., 2024) in this domain leverages the encoder-decoder architecture, which first converts images into visual representations and then translates these representations into diagnostic reports, leading to significant advancements in the field.

Despite notable progress has been achieved, generating reports with accurate clinical details remains challenging due to the inherent biases in medical imaging data. To address this, knowledge graphs (Zhang et al., 2020; Jain et al., 2021) have been introduced in the field. However, these graphs are limited by inherent challenges, such as a lack of detailed clinical knowledge or difficulties in practical application, which hinder their ability to effectively incorporate relevant clinical information into RRG. Additionally, prompting has emerged as a promising solution to this challenge. In particular, PromptMRG (Chang et al., 2024) and DKP (Bu et al., 2024b) have been introduced to integrate prompts into report generation, leading to significant improvements in the generation of clinically accurate reports. However, such success depends on extensive model parameters, which constrains its broader applicability.

To mitigate the aforementioned limitations, we introduce a radiology report generation framework called DDGIP, leveraging Disease Description Graph and Informed Prompting to effectively capture disease-specific information in report generation. First, we design a general Disease Description Graph (DDG) which comprises nodes from five key domains: **disease**, **topic**, **location**, **uncertainty**, and **severity**, encapsulating systematic and detailed clinical information. Aligning visual features with this graph can enhance the quality of the generated reports to a large extent. Next, we propose an Informed Prompting (IP) method, wherein the prompt is informed by Cross-Modal Alignment (CMA) and Bag-Of-Words (BOW) planning, thereby succeeding in RRG with a three-layer decoder and overcoming the need for extensive model parameters. The IP acts as an implicit BOW planning method, working in synergy with DDG for surface realization with clinical details. Overall, our proposed framework excels in generating high-quality reports that deliver detailed disease descriptions with improved fluency.

Our contributions can be summarized as follows:

[1] The disease description graph and model's code are available at: https://github.com/chentaohuang/DDGIP

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3884–3894
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

- We design an innovative DDG that provides clinical information for radiology report generation. This graph proves to contribute to enhancing the quality of generated reports from both quantitative and qualitative perspectives;
- We propose a novel IP method, which reduces the need for extensive model parameters in prompting while achieving significant performance with a three-layer decoder. Additionally, this method could function as an implicit BOW planning for surface realization.

2 Related work

Radiology report generation. Leveraging the latest advancements in computer vision and natural language processing, a variety of innovative approaches (Jing et al., 2017; Li et al., 2019; Chen et al., 2020; Liu et al., 2022) have emerged to integrate radiology images with free-text data, thereby enabling the automatic generation of radiology reports. The AlignTransformer (You et al., 2021) was developed to hierarchically align visual features with disease tags. Cross-modal alignment has also gained attention in the field of RRG. XPRONET(Wang et al., 2022a) established a shared subspace for effective cross-modal alignment based on prototypes. Similarly, trainable cross-modal feature matrices (Chen et al., 2022; Qin and Song, 2022; Shen et al., 2024) have been utilized to enhance alignment. Additionally, contrastive attention techniques (Liu et al., 2021b; Song et al., 2022) have achieved remarkable results by comparing target images with normal or similar images. Furthermore, advanced nonlinear attention mechanisms (Wang et al., 2022b, 2023), leveraging bilinear pooling, have been explored to capture fine-grained descriptions. (Bu et al., 2024a) proposed EKAgen for generating instance-specific expert knowledge for each query image, extracted from a knowledge support system grounded in report embeddings. Moreover, prompting methods have been introduced to advance the field of RRG. PromptMRG (Jin et al., 2024) was introduced to generate diagnosis-aware prompts based on predicted diseases, providing essential guidance for report generation. Bu et al. (2024b) introduced the Dynamic Knowledge Prompt framework to provide instance-level pulmonary lesion knowledge as prompts. Despite significant improvement, many models still struggle with omitting crucial clinical details in their generated reports.

Knowledge graph in RRG. Conveying essential clinical information is crucial for RRG. Consequently, various studies have focused on designing knowledge graphs to enhance report generation. To improve the accuracy of positive disease keywords in generated reports, Zhang et al. (2020) devised a general graph comprising 20 common abnormalities and their respective anatomical locations to aid in the generation process. Liu et al. (2021a) proposed PPKED, which leverages the preconstructed graph (Zhang et al., 2020) and retrieved reports to distill prior knowledge. Building on preconstructed graph (Zhang et al., 2020), Li et al. (2023) expanded it by adding additional nodes derived from specific retrieved RadGraph (Jain et al., 2021) triplets from similar reports, resulting in dynamic graphs. Utilizing the RadGraph (Jain et al., 2021) dataset, Yang et al. (2022) constructed a general graph and extracted triplets from the reports of similar images as specific knowledge. Huang et al. (2023) designed a general symptom graph, integrating it with visual and contextual features to incorporate clinical knowledge. Hou et al. (2023) treated disease tags as observation plans and introduced a planning-based model called ORGAN. They subsequently constructed observation graphs based on plans to facilitate tree reasoning. Although graph-based methods have achieved notable success, the existing two mainstream graphs possess the limitaions that impede their effectiveness in injecting clinical information. The general graph (Zhang et al., 2020) offers only coarse-grained knowledge. In contrast, RadGraph (Jain et al., 2021) contains over million entities, but posing a significant difficulty in utilizing this vast information to assist RRG.

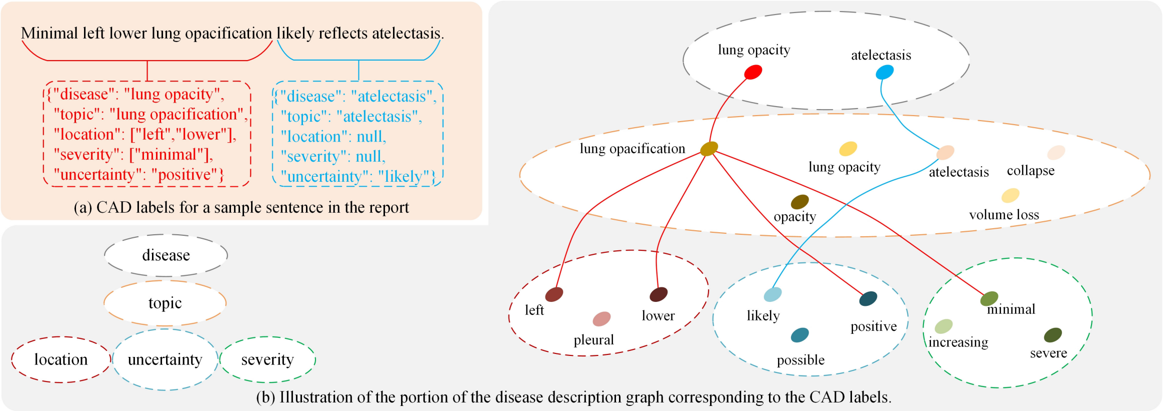
3 Method

3.1 Disease description graph

To address the limitations of existing graphs (Zhang et al., 2020; Jain et al., 2021), we design a general graph to strike a balance between knowledge comprehensiveness and generalization, offering detailed disease-specific knowledge.

The Comprehensive Annotation of Diseases (CAD)-Chest dataset[2] (Zhang et al., 2023) is derived from MIMIC-CXR radiology reports. It includes fine-grained labels about disease name, report name, severity grading, diagnostic uncertainty, and location. The report name labels encompasses

[2] https://physionet.org/content/cad-chest/1.0/



(a) CAD labels for a sample sentence in the report

(b) Illustration of the portion of the disease description graph corresponding to the CAD labels.

Figure 1: An example of the transformation from CAD labels to their corresponding part of disease description graph.

different clinical representations of the same disease. For instance, atelectasis can be referred to as collapse, volume loss, or atelectasis itself, among other variations. For better understanding, we refer to report name as topic throughout the paper.

To construct the disease description graph, these labels are first transformed into nodes that belong to five distinct domains: **disease**, **topic**, **severity**, **uncertainty**, and **location**. Then the **topic** nodes serve as anchors, and all possible co-occurrences between these anchors with other nodes are treated as edges, resulting in the disease description graph $\mathcal{G}$. The graph detailedly provides the clinical information for each disease in radiology reports in a systematic manner. In the graph, each **disease** node connects to one or more **topic** nodes, and each **topic** node links to various **severity**, **uncertainty**, and **location** nodes. The adjacency matrix of the disease description graph $\mathcal{G}$ is denoted as A. A portion of this graph with its construction process is shown in the Figure 1.

3.2 Report generation with disease description graph

Disease description graph encoding. The framework of DDGIP is depicted in Figure 2. As shown on the left side of Figure 2, we utilize a Transformer-based graph encoder (Vaswani et al., 2017) to encode the topic graph, employing the disease description graph $\mathcal{G}$ and its adjacency matrix A to derive the encoded node representation $\mathbf{N}$:

$$\mathbf{N} = Encoder_g(Embed(\mathcal{G}), A). \tag{1}$$

In this process, each node is embedded using an embedding layer $Embed$ prior to encoding, while A serves as the self-attention mask, facilitating information transfer exclusively between connected nodes.

Visual features encoding. To effectively capture the essence of the input image X, we implement a Convolutional Neural Network CNN alongside a Multilayer Perceptron MLP to extract its visual features:

$$\mathbf{X} = \{x_1, \dots, x_N\} = MLP(CNN(X)), \tag{2}$$

where x_i denotes the i-th visual feature in $\mathbf{X}$. These features are further processed by a Transformer-based visual encoder:

$$\mathbf{h^v} = Encoder_v(\mathbf{X}), \tag{3}$$

which enhances their contextual relevance and yields the visual representations $\mathbf{h^v}$.

Vision-graph alignment. Leveraging the extracted visual features $\mathbf{X}$, we align them with the graph through Cross Attention (CA), integrating the visual features with the node representations as graph attention:

$$\mathbf{h^{align}} = FFN(CA(\mathbf{X}, \mathbf{N}, \mathbf{N})). \tag{4}$$

This results in the aligned visual representation $\mathbf{h^{align}}$, merging knowledge from the visual features with the DDG.

Report generation. With the initial hidden state $\mathbf{h^1}$ as the input of the Decoder, the entire procedure of our decoder can be expressed as follows:

$$\mathbf{h^2} = MMHA(\mathbf{h^1}, \mathbf{h^1}, \mathbf{h^1}), \tag{5}$$

$$\mathbf{h^3} = CA(\mathbf{h^2}, \mathbf{h^{align}}, \mathbf{h^{align}}), \tag{6}$$

$$\mathbf{h^4} = CA(\mathbf{h^3}, \mathbf{h^v}, \mathbf{h^v}), \tag{7}$$

$$p(y_l|y_{1:l-1}, X) = Softmax(\mathbf{W}\mathbf{h}_l^4 + \mathbf{b}), \tag{8}$$

where $MMHA$ represents the masked multi-head self-attention and $\mathbf{h^1}$,$\mathbf{h^2}$,$\mathbf{h^3}$,$\mathbf{h^4}$ are temporary hidden states in the decoding phase. $\mathbf{W} \in \mathbb{R}^{|V| \times d}$ is the weight matrix, and $\mathbf{b} \in \mathbb{R}^{|V|}$ is the bias vector, with $|V|$ representing the vocabulary size and d representing the model size. For the sake of simplicity, we omit modules from the standard Transformer such as Layer Normalization and Feed-Forward Network.

3.3 Informed prompting

Visual prompting (Wu et al., 2024) has been widely adopted in large vision-language models, achieving notable success across various tasks. However, its effectiveness often relies on large-scale data and extensive model parameters, which limits its applicability in RRG. Generally, visual prompting aims at optimizing the objective $P(Y|P,X)$ where Y represents the text, X represents the image, and P represents the prompt derived from X. However, prompting can be seen as the process of planning the words to be generated, which suggests that textual data may facilitate the prompt-learning process more efficiently than visual data. Therefore, we introduce a variational inference strategy to learn the prompt using textual data while aligning the visual modality with the textual one. This approach serves as a "shortcut" enabling efficient prompt learning without the need for large datasets and extensive model parameters. Besides, we introduce the BOW planning task to enrich the prompt with word distribution information for surface realization. This informed prompt is designed to synergize with the DDG, enabling the generation of reports with precise disease descriptions. In conclusion, we propose a novel and efficient method: informed prompting, leveraging cross-modal alignment (Najdenkoska et al., 2022) and bag-of-words planning (Fu et al., 2019; Hu et al., 2022).

Cross-modal alignment. We treat RRG as a process of generating a report Y for the given image X. From a probabilistic perspective, the goal of optimization in the process is to maximizing the conditional log-likelihood: $logP(Y|X)$. Next, we incorporate the latent variable θ into the conditional log-likelihood, we have:

$$\begin{aligned} logp(Y|X) = &\int q(\theta) log \frac{p(Y,\theta|X)}{q(\theta)} \, d(\theta) \\ &+ D_{KL}(q(\theta)||p(\theta|Y,X)), \end{aligned} \tag{9}$$

where $q(\theta)$ represents the variational distribution to approximate the posterior distribution $p(\theta|Y,X)$. By minimizing the KL divergence term, $D_{KL}(q(\theta)||p(\theta|Y,X))$, we get the ELBO of the log-likelihood:

$$\begin{aligned} logp(Y|X) \geq &\int q(\theta) log \frac{p(Y,\theta|X)}{q(\theta)} \, d(\theta) \\ &= ELBO, \end{aligned} \tag{10}$$

which can be formulated as:

$$\begin{aligned} ELBO = &\mathbb{E}(logp(Y|\theta,X)) \\ &- D_{KL}(q(\theta)||p(\theta|X)). \end{aligned} \tag{11}$$

Then we incorporate the text modality into the training process by conditioning the variational distribution $q(\theta)$ on the ground-truth report Y, as $q(\theta|Y)$. Based on the ELBO, we derive the objective function with respect to a report Y as follows:

$$\begin{aligned} \mathcal{L}_{ELBO} = &-\sum_{l=1}^{L}[\frac{1}{K}\sum_{k=1}^{K} logp(y_l|y_{1:l-1},\theta^{\{k\}},X)] \\ &+ \alpha D_{KL}[q(\theta|Y)||p(\theta|X)], \end{aligned} \tag{12}$$

where $\theta^{\{k\}}$ denotes the k-th of K Monte Carlo samples, while α acts as a hyperparameter for controlling its impact in training.

For the sake of computational efficiency, we compute the average of the K samples of latent variable as $\overline{\theta}$, which is then employed to assist generation. As a result, the objective function is simplified to:

$$\begin{aligned} \mathcal{L}_{ELBO} = &-\sum_{l=1}^{L}[logp(y_l|y_{1:l-1},\overline{\theta},X)] \\ &+ \alpha D_{KL}[q(\theta|Y)||p(\theta|X)], \end{aligned} \tag{13}$$

where the first term represents the negative conditional log-likelihood loss, while the second term accounts for the KL divergence between the variational distribution $q(\theta|Y)$ and the true distribution $p(\theta|X)$, utilized to align the image modality with the text modality. During training, the samples are drawn from $q(\theta|Y)$, whereas during inference, the samples are drawn from $p(\theta|X)$.

Prompt deriving. Before introduce the informed prompt, we parameterize $q(\theta|Y)$ and $p(\theta|X)$ as:

$$q(\theta|Y) = N(\theta|h_g^t, \sigma_t^2 I), \tag{14}$$

$$p(\theta|X) = N(\theta|h_g^v, \sigma_v^2 I), \tag{15}$$

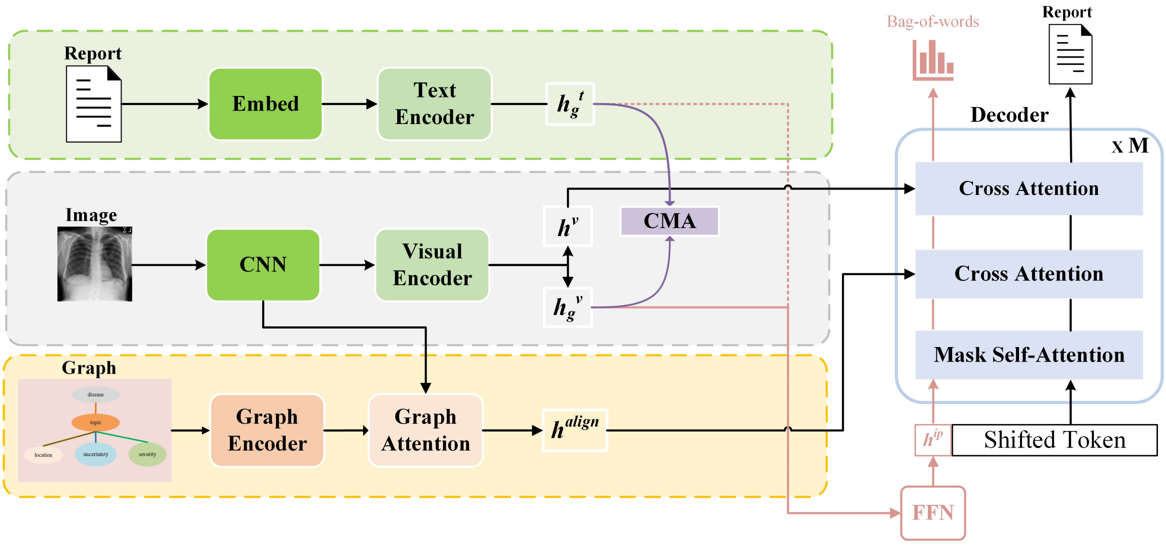


Figure 2: The overall framework of DDGIP, where CMA refers to the cross-modal alignment. The dotted line indicates the global text representation h_g^t is only available in training stage.

where σ_t and σ_v symbolize the standard deviations as learnable parameters, and h_g^t represents the global textual representation while h_g^v represents the global visual representation.

To obtain h_g^t, we transforme ground-truth report Y into textual features through a embedding process:

$$\mathbf{Y} = \{y_1, \ldots, y_L\} = Embed(Y) + P, \qquad (16)$$

where y_l represents the l-th textual feature in $\mathbf{Y}$ and P represents the positional embeddings, adding positional context to each word. These embedded textual features are then encoded using a Transformer-based text encoder to obtain the textual representations $\mathbf{h^t}$:

$$\mathbf{h^t} = \{h_1^t, \ldots, h_L^t\} = Encoder_t(\mathbf{Y}), \qquad (17)$$

with the global textual representation h_g^t being distilled into the mean value of $\mathbf{h^t}$.

To derive h_g^v, we revisit the visual features encoding process. Starting with the visual features $\mathbf{X}$, the global visual feature x_g is obtained by averaging $\mathbf{X}$. The encoding process then proceeds as follows:

$$[h_g^v, \mathbf{h^v}] = Encoder_v([x_g, \mathbf{X}]). \qquad (18)$$

To get the informed prompt, we samples θ using the reparameterization trick. Given the mean value $\overline{\theta}$ of the latent variable samples, the informed prompt h^{ip} is derived by processing $\overline{\theta}$ through a Feed-Forward Network FFN, expressed as:

$$h^{ip} = FFN(\overline{\theta}). \qquad (19)$$

The derived informed prompt is then concatenated with shifted tokens. In generation, each token attends the previously generated words and the informed prompt as guidance.

During training, samples are drawn from $q(\theta|Y)$, whereas during inference, they are drawn from $p(\theta|X)$. The distribution $p(\theta|X)$ is optimized to align with $q(\theta|Y)$ by minimizing the KL divergence, ensuring that the prompt is effectively informed through cross-modal alignment.

Bag-of-words planning task informing. Here we introduce the BOW planning task into the learning process to ensure that the prompt h^{ip} is informed by word distribution information. The BOW of the report Y is defined as a categorical distribution across the entire vocabulary:

$$P(Y|h^{ip}) = softmax(MLP(Decoder(h^{ip}))), \qquad (20)$$

where $MLP : \mathbb{R}^d \to \mathbb{R}^{|V|}$ denotes a multi-layer perceptron and $Decoder$ represents our decoder. The objective of the task is to maximize the log-likelihood of correctly predicting the occurrence of each word in the report:

$$\mathcal{L}_{BOW} = -\frac{1}{L}\sum_{l=1}^{L} logp(y_l|h^{ip}), \qquad (21)$$

where $p(y_l|h^{ip})$ represents the estimated probability of the l-th word in the BOW. From this standpoint, the informed prompting serves as an implicit BOW planning method for surface realization.

We holistically optimize our model by combining the objective function $\mathcal{L}_{ELBO}$ with the BOW planning loss $\mathcal{L}_{BOW}$:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \beta \mathcal{L}_{BOW}, \tag{22}$$

where β govern the impact of the BOW planning, acting as hyper-parameters.

4 Experiments

4.1 Datasets

We conduct experiments on two widely-used medical report generation datasets (i.e., IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019)) to evaluate our proposed model. For a fair comparison, we adopt the settings in **R2GEN** (Chen et al., 2020) to preprocess the reports.

IU-Xray[3] is the most widely-used public Chest X-ray dataset provided by Indiana University, containing 3,955 radiology reports and 7,470 chest X-ray images. Following the previous research (Chen et al., 2020), we split the data into training, validation, and testing sets with a ratio of 7:1:2.

MIMIC-CXR[4] is the largest public dataset in report generation provided by the Beth Israel Deaconess Medical Center, which includes 377,110 chest X-ray images and 227,827 reports. We adopt the official train/validation/test splits.

Given that the reports in IU-Xray are considerably simpler than those in MIMIC-CXR, for IU-Xray we begin by filtering out nodes and their corresponding edges that are not present in the reports, subsequently constructing the disease description graph. The disease description graph for IU-Xray consists of 109 nodes—20 for **disease**, 40 for **topic**, 12 for **uncertainty**, 23 for **location**, and 14 for **severity**—totaling 1,349 edges. In contrast, the disease description graph for MIMIC-CXR comprises 156 nodes—26 for **disease**, 58 for **topic**, 30 for **uncertainty**, 24 for **location**, and 18 for **severity**—and includes a total of 2,116 edges.

4.2 Evaluation metrics

To evaluate our model, we employ a comprehensive set of metrics derived from the domains of natural language generation (NLG) and clinical efficacy (CE). The adopted NLG metrics include widely recognized captioning metrics such as BLEU-n (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004), following the standard evaluation protocols. For the CE metrics, we use CheXpert (Irvin et al., 2019) to annotate the generated reports from the MIMIC-CXR dataset and compare them against the disease labels in the ground-truth reports.

4.3 Experimental setup

We utilize the ResNet-101 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as the visual feature extractor. Each encoder and decoder is configured with 3 layers, while the Graph Encoder is designed with 2 layers. Each multi-head attention block contains 8 attention heads. Our model dimensionality d is set to 512.

For training, we conduct 15 epochs for the IU-Xray dataset and 5 epochs for the MIMIC-CXR dataset, with batch sizes of 32 and 64, respectively. The best checkpoints are selected based on the BLEU-4 on the validation set. The learning rates are set to 5e-5 for the visual extractor and 1e-4 for other parameters. We decay the learning rate by a factor of 0.9/0.8 per epoch for each dataset and set the beam size to 3. The parameters α and β are set to 0.0008 and 0.05 respectively. The standard deviations σ_t and σ_v are initialized to 0.1. We use monotonic annealing schedule (Bowman et al., 2015) to learn the KL divergence term of loss. Sampling number K is set to 7. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU.

4.4 Quantitative results

To demonstrate the effectiveness of our model, we compare DDGIP against several state-of-the-art methods. Table 1 presents the outcomes of NLG and CE metrics on the IU-Xray and MIMIC-CXR datasets. The results demonstrate that DDGIP surpasses other baseline models, establishing a new state-of-the-art benchmark and showcasing the effectiveness and generalizability of our approach. Specifically, in terms of natural language generation, DDGIP achieves the highest scores on BLEU-2 and BLEU-3 across both datasets, highlighting its superiority in short-gram predictions. Moreover, our BLEU-1 score of 0.433 outperforms the second-best result by a significant 1.4% margin on MIMIC-CXR, indicating that our model excels in generating reports with precise terminology. Additionally, DDGIP achieves the top METEOR score of 0.167, further solidifying its effectiveness. These advancements could contribute to the collab-

[3] https://openi.nlm.nih.gov/

[4] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

Dataset	Model	NLG metrics						CE metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	Precision	Recall	F1-score
IU-XRAY	R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	-	0.371	-	-	-
	PPKED (Liu et al., 2021a)	0.483	0.315	0.224	0.168	-	0.376	-	-	-
	KnowMat (Yang et al., 2022)	0.496	0.327	0.238	0.178	-	0.381	-	-	-
	DCL (Song et al., 2022)	-	-	-	0.163	0.193	0.383	-	-	-
	METrans (Wang et al., 2023)	0.483	0.322	0.228	0.172	0.192	0.380	-	-	-
	KIUT (Huang et al., 2023)	0.525	0.360	0.251	0.185	**0.242**	0.409	-	-	-
	ORGAN (Huang et al., 2023)	0.510	0.346	0.255	0.195	0.205	0.399	-	-	-
	PromptMRG (Jin et al., 2024)	0.401	-	-	0.098	0.160	0.281	-	-	-
	EKAgen (Bu et al., 2024a)	**0.526**	0.361	0.267	**0.203**	0.214	0.404	-	-	-
	DKP (Bu et al., 2024b)	0.507	0.344	0.245	0.181	0.214	0.398	-	-	-
	DDGIP	0.513	**0.365**	**0.269**	0.202	0.226	**0.411**	-	-	-
MIMIC-CXR	R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	PPKED (Liu et al., 2021a)	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	KnowMat (Yang et al., 2022)	0.363	0.228	0.156	0.115	-	0.284	0.458	0.348	0.371
	DCL (Song et al., 2022)	-	-	-	0.109	0.150	0.284	0.471	0.352	0.373
	METrans (Wang et al., 2023)	0.386	0.250	0.169	**0.124**	0.152	0.291	0.364	0.309	0.311
	KIUT (Huang et al., 2023)	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	ORGAN (Huang et al., 2023)	0.407	0.256	**0.172**	0.123	0.162	**0.293**	0.416	0.418	0.385
	PromptMRG (Jin et al., 2024)	0.398	-	-	0.112	0.157	0.268	0.501	0.509	0.476
	EKAgen (Bu et al., 2024a)	0.419	0.258	0.170	0.119	0.157	0.287	**0.517**	0.483	0.499
	DKP (Bu et al., 2024b)	0.418	0.260	**0.172**	0.120	0.159	0.291	0.496	0.461	0.478
	DDGIP	**0.433**	**0.266**	**0.172**	0.117	**0.167**	0.288	0.508	**0.529**	**0.518**

Table 1: Comparison with SOTA methods on the IU-XRAY and MIMIC-CXR datasets with NLG and CE metrics. The best results are in **bold**. The second-best results are underlined. - indicates that the metric was not measured in the cited paper.

oration between DDG and IP to generate detailed descriptions of disease. Notably, DDGIP outperforms other graph-based models (**PPKED**, **KnowMat**, **DCL**, **KIUT**, **ORGAN**) as well as prompt-based models (**PromptMRG**, **DKP**).

In terms of clinical efficacy, our model also outperforms other baselines. Notably, DDGIP achieves the highest Recall score of 0.529, leading the second-best result by a 2.0% improvement, demonstrating its superior ability to diagnose abnormalities. However, a gap remains in the Precision score between DDGIP and the top-performing model, **EKAgen**. In future work, incorporating planning on the **topic** nodes of DDG could help mitigate this gap and further enhance the quality of the generated reports.

4.5 Ablation study

Analysis of DDGIP. We perform ablation studies to elucidate the contributions of our proposed DDG and IP. The results are presented in Table 2. For our base model, we utilize a vanilla Transformer. A comparison between the Base and the Base+DDG reveals an overall improvement, particularly on BLEU-n scores, indicating that graph-aligned visual features improves the accuracy of generated content, boosting its quality. When we analyze the Base+IP, we observe substantial advancements in BLEU-1 and BLEU-2 scores, demonstrating that IP effectively guides the generation of short grams. Improvemnts in BLEU-4 and METEOR scores suggest that IP also enhances the fluency of the generated reports. Furthermore, the combination of DDG and IP showcases their synergistic effect, propelling DDGIP to achieve state-of-the-art results.

Analysis of informed prompting. Recognizing the crucial impact of informed prompting, we conduct a thorough analysis of its performance. Initially, we examine the contributions of CMA and BOW planning within the informed prompting. As indicated in Table 3, the most significant advancements arise from CMA, while BOW planning yields modest overall improvements. The effectiveness of CMA confirms that the variational inference strategy can serve as a "shortcut" for prompt learning, significantly reducing the need for large datasets and extensive parameters. Notably, when comparing model (a) in Table 3 with the Base in Table 2, we find that merely integrating a non-informed prompt to the decoder could result in slight development. Next, we explore the effects of informed prompting across different decoder designs. The results shown in Table 4 illustrate that informed prompting in the decoder-only structure leads to generate accurate short-gram out-

Dataset	Model	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU-Xray	Base	0.393	0.258	0.189	0.146	0.172	0.362
	Base+DDG	0.456	0.310	0.231	0.182	0.194	0.385
	Base+IP	**0.515**	0.359	0.260	0.198	0.221	0.407
	DDGIP	0.513	**0.365**	**0.269**	**0.202**	**0.226**	**0.411**
MIMIC-CXR	Base	0.314	0.198	0.136	0.099	0.130	0.275
	Base+DDG	0.336	0.211	0.145	0.107	0.138	0.280
	Base+IP	0.420	0.255	0.164	0.111	0.163	0.282
	DDGIP	**0.433**	**0.266**	**0.172**	**0.117**	**0.167**	**0.288**

Table 2: Ablation study of DDGIP, where DDG refers to the Disease Description Graph and IP refers to the Informed Prompting. The best results are in **bold**.

Dataset	Model	CMA	BOW Planing	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU-Xray	(a)	✗	✗	0.401	0.264	0.192	0.148	0.174	0.370
	(b)	✗	✓	0.420	0.276	0.200	0.153	0.180	0.372
	(c)	✓	✗	0.507	0.347	0.250	0.194	0.218	0.403
	(d)	✓	✓	**0.515**	**0.359**	**0.260**	**0.198**	**0.221**	**0.407**
MIMIC-CXR	(a)	✗	✗	0.317	0.199	0.137	0.101	0.131	0.275
	(b)	✗	✓	0.321	0.202	0.139	0.103	0.132	0.276
	(c)	✓	✗	0.418	0.254	0.163	0.110	0.161	0.281
	(d)	✓	✓	**0.420**	**0.255**	**0.164**	**0.111**	**0.163**	**0.282**

Table 3: Ablation study of informed prompting. In experiments, models without CMA sample θ from $p(\theta|X)$ during both the training and inference stages. The best results are in **bold**.

Dataset	Decoder	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU-Xray	A	0.393	0.258	0.189	0.146	0.172	0.362
	B	0.483	0.334	0.243	0.182	0.208	0.403
	C	**0.515**	**0.359**	**0.260**	**0.198**	**0.221**	**0.407**
MIMIC-CXR	A	0.314	0.198	0.136	0.099	0.130	0.275
	B	0.404	0.233	0.143	0.093	0.158	0.266
	C	**0.420**	**0.255**	**0.164**	**0.111**	**0.163**	**0.282**

Table 4: Ablation study of decoder designs. In this study, Decoder A represents the vanilla transformer decoder, Decoder B refers to the decoder-only structure which excludes Cross Attention but incorporates the informed prompt, and Decoder C signifies the vanilla transformer decoder incorporated with the informed prompt. The best results are in **bold**.

Graph variants	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
DDG *w/o* **disease**	0.415	0.277	0.205	0.162	0.178	0.376
DDG *w/o* **location**	0.410	0.273	0.201	0.159	0.177	0.371
DDG *w/o* **uncertainty**	0.435	0.285	0.203	0.153	0.179	0.367
DDG *w/o* **severity**	0.423	0.279	0.205	0.158	0.180	0.370
DDG	**0.456**	**0.310**	**0.231**	**0.182**	**0.194**	**0.385**

Table 5: Ablation study of the domain nodes in DDG on the IU-Xray dataset. The best results are in **bold**.

puts on both datasets, although it exhibits weaker performance on BLEU-4 for MIMIC-CXR, which is a larger and more complex dataset. However, by integrating informed prompting into a vanilla Transformer decoder, we have successfully addressed this limitation, boosting the BLEU-4 score on MIMIC-CXR from 0.093 to 0.111.

Edge	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
Undirected	0.455	0.306	0.225	0.173	0.192	0.380
Directed	**0.456**	**0.310**	**0.231**	**0.182**	**0.194**	**0.385**

Table 6: Ablation study of the edge connections in DDG on the IU-Xray dataset. The best results are in **bold**.

Analysis of disease description graph. To explore the impact of the disease description graph, we conduct ablation studies to evaluate the effect of each individual component. The results for the graph variants, presented in Table 5, reveal that each domain of the DDG plays a pivotal role in enhancing performance, with the domain working synergistically to improve the model's efficacy. Additionally, the results for edge connections, shown in Table 6, indicate that directed edges outperform undirected ones. This can be explained by the fact that undirected edges connect irrelevant **topic** nodes through **uncertainty**, **severity**, and **location** nodes, reducing the distinction between **topic** nodes and introducing challenges in training.

4.6 Qualitative analysis

To intuitively illustrate the effectiveness of the proposed DDG and IP, we present a qualitative example in Figure 3. As shown, our model, assisted by the disease description graph, generates reports

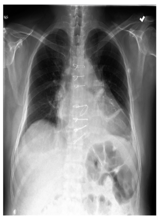
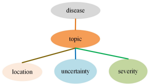
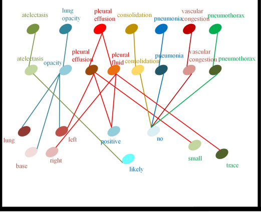
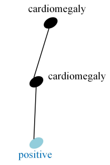
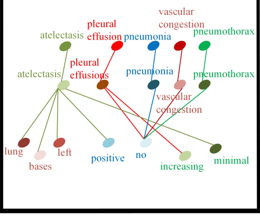
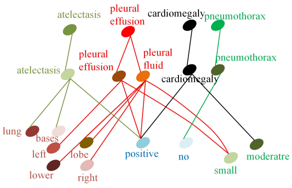
Image	Ground Truth	Base	Base+DDG	DDGIP
	there is persistent opacity at the left lung base though aeration is improved from two days prior . persistent linear opacities likely represent atelectasis . left pleural effusion is small . trace right pleural fluid is also present . lungs are otherwise well aerated . there is no focal consolidation to suggest pneumonia . there is no vascular congestion or pulmonary edema . there is no pneumothorax .	as compared to the previous radiograph there is no relevant change . status post sternotomy . the alignment of the sternal wires is constant . the alignment of the sternal wires is constant . constant appearance of the cardiac silhouette .	as compared to the previous radiograph the left chest tube has been removed . there is no evidence of pneumothorax . minimal atelectasis at the left lung bases . no larger pleural effusions . no pulmonary edema . no pneumonia .	as compared to the previous radiograph the patient has been extubated . there is a small amount of left pleural fluid in the right lower lobe and atelectasis . moderate cardiomegaly is unchanged . median sternotomy wires are intact . interval removal of the previously described small left pleural effusion . there is no evidence of pneumothorax . atelectasis at the lung bases .
Graph	**Ground Truth**	**Base**	**Base+DDG**	**DDGIP**
				

Figure 3: A qualitative example showcasing our model, along with the generated reports and their corresponding representations in the disease description graph.

that are not only more clinically accurate but also pertinent to the details of diseases, featuring descriptions such as "minimal" and "left lung bases". Furthermore, with the aid of the informed prompt, the model accurately generates phrases like "a small amount of left pleural fluid in the right lower lobe" and "small left pleural effusion", significantly enhancing the quality of the generated reports.

However, despite these successes, a notable error must be addressed. Topics such as "consolidation" and "cardiomegaly" are either omitted or incorrectly generated in the report. This highlights the need for better alignment between visual features and the graph. In future work, incorporating planning on the **topic** nodes of DDG is expected to address this challenge.

5 Conclusion

In this study, we present an innovative framework DDGIP for radiology report generation which leverages the disease description graph and informed prompting. The disease description graph is designed to inject detailed disease-specific knowledge into report generation. Additionally, we propose an informed prompting method to guide the generation, enhancing the precision and fluency of generated content. The synergistic effect between the disease description graph and informed prompting results in a significant improvement in the quality of the generated reports. Experimental results on the IU-Xray and MIMIC-CXR datasets demonstrate that our approach achieves state-of-the-art performance.

Limitation. The primary limitation of our framework is the occasional omission or inaccuracy in generating certain clinical topic descriptions, which may affect the clinical efficacy. Looking ahead, we aim to integrate planning on the **topic** nodes of DDG, thereby enhancing vision-graph alignment and boosting the model's clinical efficacy to address this challenge.

Ethics Statement

The IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets have undergone automatic de-identification to ensure patient privacy. The proposed system aims to automate the generation of radiology reports, thereby alleviating the workload for radiologists. However, we have observed instances where the system may inaccurately generate or omit disease descriptions, leading to potential diagnostic errors. If the system learns from additional private datasets after deployment, there are risks of personal information leakage through the generated reports. To mitigate these risks and enhance privacy protection, further anonymization technologies should be implemented. Therefore, we encourage users to carefully consider the ethical implications of the generated outputs in real-world applications.

Acknowledgments

This research was partly funded by the National Natural Science Foundation of China (Grant Nos. 62361027 and 62161011), the Key Research and Development Plan of Jiangxi Provincial Science and Technology Department (Key Project) (Grant No. 20223BBE51036), the Humanity and Social Science Fund of Ministry of Education of China (Grant No. 23YJA870005), the Natural Science Foundation of Jiangxi Provincial Department of Science and Technology (Grant Nos. 20232BAB202022 and 20232BAB202004), the Humanity and Social Science Foundation of the Jiangxi Province (Grant No. 22TQ01), the Graduate Innovation Foundation Project of Jiangxi Province (Grant No. YC2023-S477).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 2024a. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204.

Shenshen Bu, Yujie Song, Taiji Li, and Zhiming Dai. 2024b. Dynamic knowledge prompt for chest x-ray report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5425–5436.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.

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

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. *Advances in Neural Information Processing Systems*, 32.

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.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. Organ: observation-guided radiology report generation via tree reasoning. *arXiv preprint arXiv:2306.06466*.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. Planet: Dynamic content planning in autoregressive transformers for long-form text generation. *arXiv preprint arXiv:2203.09100*.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. 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.

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*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

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 2019. *arXiv preprint arXiv:1901.07042*.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6666–6673.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

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

Fenglin Liu, Shen Ge, Yuexian Zou, and Xian Wu. 2022. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. 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, Yuexian Zou, Ping Zhang, and Xu Sun. 2021b. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. 2022. Uncertainty-aware report generation for chest x-rays by variational topic inference. *Medical Image Analysis*, 82:102603.

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.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4776–4783.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. 2022. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2388–2397.

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.

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

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.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022b. A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

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*, 80:102510.

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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer.

Mengliang Zhang, Xinyue Hu, Lin Gu, Tatsuya Harada, Kazuma Kobayashi, Ronald Summers, and Yingying Zhu. 2023. Cad-chest: Comprehensive annotation of diseases based on mimic-cxr radiology report.

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*, volume 34, pages 12910–12917.