Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20206-x

DMR^{2}G: diffusion model for radiology report generation

Huan Ouyang [1] · **Zheng Chang** [1] · **Binghao Tang** [1] · **Si Li** [1]

Received: 2 June 2024 / Revised: 10 August 2024 / Accepted: 29 August 2024

Abstract
Radiology report generation aims to generate pathological assessments from given radiographic images accurately. Prior methods largely rely on autoregressive models, where the sequential token-by-token generation process always results in longer inference time and suffers from the sequential error accumulation. In order to enhance the efficiency of report generation without compromising diagnostic accuracy, we present a novel radiology report generation approach based on diffusion models. By integrating a graph-guided image feature extractor informed by a radiology knowledge graph, our model adeptly identifies critical abnormalities within images. We also introduce an auxiliary lesion classification loss mechanism using pseudo labels as supervision to align image features and textual disease keyword representations accurately. By adopting the accelerated sampling strategy inherent to diffusion models, our approach significantly reduces the inference time. Through comprehensive evaluation on the IU-Xray and MIMIC-CXR benchmarks, our approach outperforms autoregressive models in inference speed while maintaining high quality, offering a significant advancement in automating radiology report generation task.

Keywords Radiology report generation · Diffusion model · Lesion feature extraction · Medical knowledge graph

1 Introduction

In contemporary clinical settings, radiology images are pivotal in diagnosing many conditions, such as pneumonia and pneumothorax. Typically, interpreting these images and the subsequent composition of examination reports demands a substantial investment of time

Huan Ouyang and Zheng Chang contributed equally to this work.

✉ Si Li
lisi@bupt.edu.cn

Huan Ouyang
ouyanghuan@bupt.edu.cn

Zheng Chang
zhengchang98@bupt.edu.cn

Binghao Tang
tangbinghao@bupt.edu.cn

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

from skilled medical practitioners, imposing a considerable workload. The automation of radiology report generation emerges as a promising solution to mitigate this challenge. Consequently, this task has garnered significant attention from the research community in recent years [1–6].

Early works in radiology report generation primarily extend from image captioning techniques, which produce textual descriptions conditioned on the given image. Notably, [7] leveraged CNNs for feature extraction and HRNNs [8] for report generation, while CMAS [4] employed dual RNNs to enhance sentence accuracy in reports. However, radiology report generation presents unique challenges compared to image captioning due to significant biases in visual and textual data [2, 9]. Radiology images pose difficulties in highlighting small, critical areas like lesions. This necessitates incorporating prior knowledge for better lesion detection. To address this, Zhang et al. [1] introduced a knowledge graph to assist in finding relationships among chest normal or disease keywords and emphasize disease terms. Moreover, recent works [10–12] have applied contrastive learning to enhance representations of visual abnormal regions and textual disease keywords, which have departed from traditional image captioning approaches.

However, all these works are based on autoregressive models, which generate sentences token by token suffering from unsatisfactory inference speed as illustrated in the top part of Fig. 1 and sequential error accumulation [13]. To address the issues, researchers have begun to explore the utilization of non-autoregressive (NAR) architectures in the text generation task. NAT [14] firstly proposes the non-autoregressive transformer in the machine translation task and generates the entire output sequence in parallel, thereby improving generation speed and efficiency. COLD [15] proposed a decoding method that relaxes the discrete language model outputs to continuous variables and backpropagates gradient information from the right context. However, these NAR methods exhibit limited scalability and pose challenges in integrating prior knowledge, with evidence suggesting their inadequacy for comprehensive language modeling [16].

Diffusion model [17] is impressive in their generative quality and versatility in image generation. In text generation, it enables the parallel generation of all tokens as shown in the

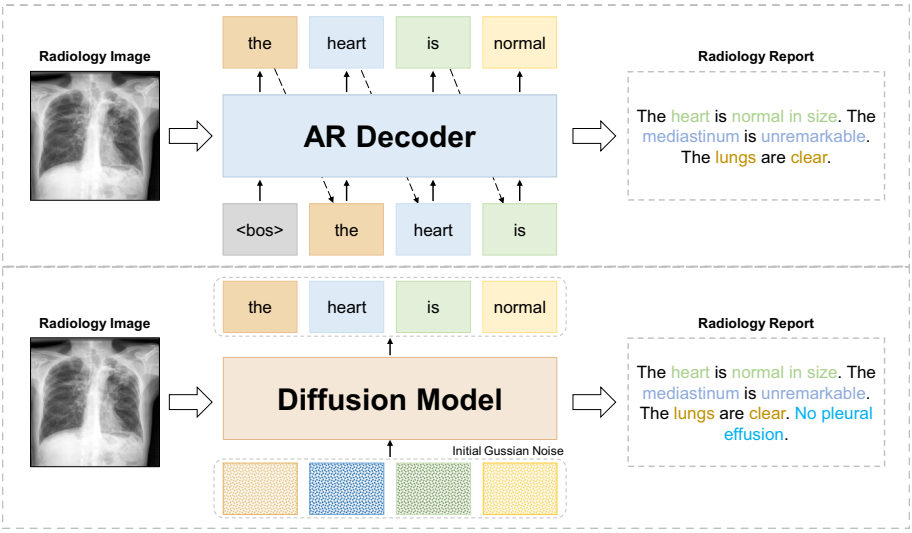

Fig. 1 Difference of the autoregressive decoder and diffusion model in radiology report generation

bottom part of Fig. 1. Some progress has been made in improving diffusion models for text generation. Diffusion-LM [18] and Diffuseq [19] take continuous text representations (e.g., word embeddings or hidden states) as training targets and conduct diffusion processes in the corresponding latent space. Besides, Diffusion models also exhibit excellent scalability to facilitate the integration of cross-modal information. Liu et al. [20] connects image space and text space by using a lightweight mapping network as prior knowledge. Furthermore, in the generative process, employing sampling methods at timesteps can significantly accelerate the inference speed. Thus, exploring the application of diffusion models in radiology report generation is of value.

In this paper, we propose a radiology report generation method based on the diffusion model (DMR^{2}G), breaking the previous paradigm that relied on the autoregression framework for the radiology report generation task. We introduce prior knowledge to augment the quality of generated reports and employ the DDIM [21] method to reduce sampling timesteps, thereby expediting the inference speed. Leveraging the scalability of the diffusion model, we design the graph-guided image feature extractor, which is grounded in the general radiology knowledge graph. This extractor utilizes the prior knowledge embedded within the pre-construct graph to aid in the identification of abnormal regions within the image. We also propose the auxiliary lesion classification loss as a supervisory signal, employing pseudo labels to align image abnormal region features and textual disease keywords. We evaluate our method on two benchmarks, IU-Xray [22] and MIMIC-CXR [23]. During the inference process, our model outperforms most autoregressive-based models in inference speed while maintaining comparable report generation quality. In summary, our main contributions are as follows:

- We propose an innovative method based on the diffusion model for radiology report generation, exploring a new framework for this task.
- We design the graph-guided feature extractor to integrate prior knowledge and utilize the auxiliary lesion classification loss to enhance the visual-textual alignment.
- We apply accelerated sampling strategies to greatly accelerate the inference speed while maintaining the accuracy of report generation at a comparable level.

2 Related work

2.1 Image captioning

Image captioning aims to generate human-like sentences to describe a given image. This task [24, 25] is considered a high-level visual understanding problem that combines the research of computer vision and natural language processing. Early models [25–28] use a visual encoder to extract visual features and apply recurrent neural network as decoder for caption generation. Later on, attention-based methods have been proposed to capture multimodal alignment [29–31] and perform object-relational reasoning [32, 33]. Besides, researchers have explored to leverage semantic attributes [34–37] and scene graphs [38] for captioning. These approaches mainly adopt the encoder-decoder framework and have demonstrated a great improvement in some traditional image captioning benchmarks. However, rather than only generating one single sentence, radiology report generation aims to generate a long paragraph, which consists of multiple structural sentences with each one focusing on a specific medical observation for a specific region in the radiology image.

2.2 Radiology report generation

Writing a radiology report can be time-consuming and tedious for experienced radiologists, and error-prone for un-experienced radiologists. Most previous works [3, 4, 7] attempt to adopt an HRNN to automatically generate a fluent report. However, due to the serious data deviation, these models are poor at finding visual groundings and are biased towards generating plausible but general reports without prominent abnormal narratives. Recently, some approaches [1, 2, 5, 6] have been proposed to alleviate data deviation. HRGR-Agent [2] proposed a hybrid model using template retrieval and text generation, focusing on the generated normal and abnormal sentences to enhance the model's ability to describe abnormalities. Additionally, ASGK [9] introduces the medical graph to enhance the understanding of the relationships among lesions. Additionally, with the rise of large language models, researchers have explored their integration into radiology report generation. Notable works, [39, 40] have demonstrated that these models help mitigate the problems associated with data deviation and can generate more nuanced and informative narratives. Consequently, we also design corresponding modules to incorporate medical prior knowledge.

2.3 Diffusion model for text generation

Existing generative models like GAN [41] and VAE [42] have problems such as training instability, mode collapsing. While solving these problems, diffusion models have state-of-the-art sample quality in many tasks [43–45]. DDPM [17] proposed a parameterized Markov chain trained by variational inference. However it is not the diffusion model's nature to deal with discrete data, some works have been proposed to tackle this problem, and ARDMs [46] introduced a new model at the intersection of autoregressive models and discrete diffusion models. Analog bits [47] represented the discrete data as binary bits and used a continuous diffusion model to model these bits. Diffusion-LM [18] transferred tokens to continuous embedding representations and modeled them with continuous diffusion models. Furthermore, DiffCap [48] explores a continuous diffusion model on image captioning tasks. The diffusion model offers the benefits of consistent training stability and high-quality generation in many tasks. Therefore, we further explore the new framework based on continuous diffusion model for radiology report generation.

3 Methodology

In this section, we provide a detailed introduction to the proposed diffusion-based framework for radiology report generation shown in Fig. 2. To make this paper self-contained, we first offer an overview of diffusion model and the method for accelerated sampling to improve the model's inference speed in Section 3.1. In Section 3.2, we introduce the proposed framework to use the diffusion model for radiology report generation. Then we describe the graph-guided image feature extractor in Section 3.3, which is utilized to integrate pathological prior knowledge. Finally, we present the abnormality classification loss to enhance the alignment between pathological region features in images and lesion entity keywords in Section 3.4.

3.1 Preliminaries

Diffusion model The diffusion model is a generative model that simulates the process of gradually adding noise to data, and then learning to reverse this process to generate new samples.

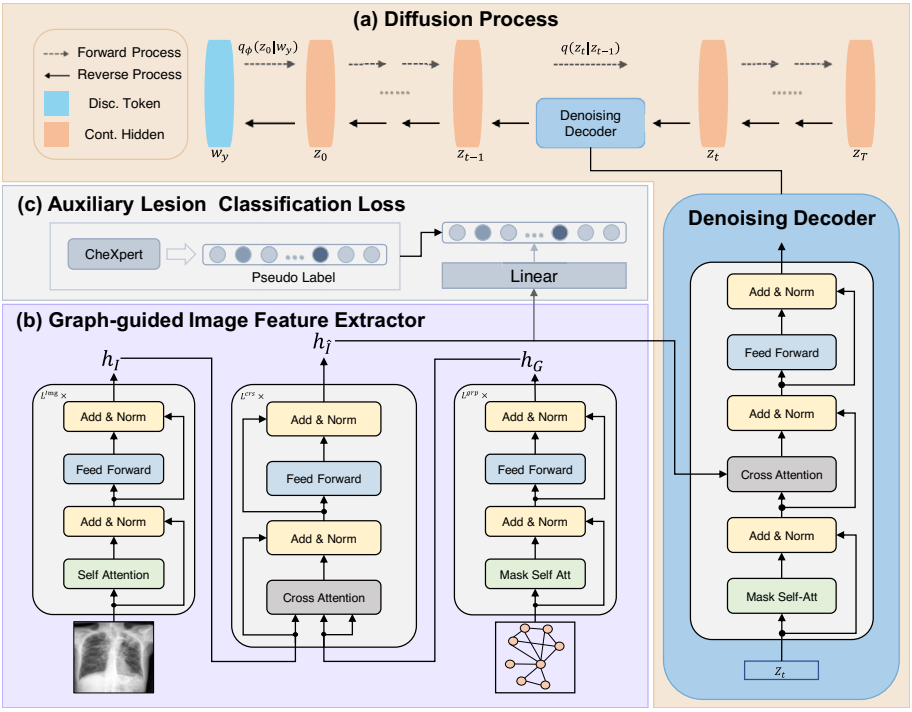

Fig. 2 The framework of DMR^{2}G. (a) We propose a radiology report generation method based on the diffusion model, breaking the previous paradigm that relied on the autoregression framework (Section 3.2). During the diffusion process, we employ the denoising decoder to facilitate the interaction between the predicted noise and the extracted image features to generate the next report embedding. (b) We design the graph-guided image feature extractor, which is grounded in the general radiology knowledge graph (Section 3.3). (c) We also propose the auxiliary lesion classification loss as a supervisory signal, employing pseudo labels to align image abnormal region features and textual disease keywords (Section 3.4)

A Markov chain mathematically describes the forward process, where each step adds a small amount of Gaussian noise, following a predefined schedule. At each time step $t \in \{1, 2, \ldots, T\}$, a noise sample z_t is sampled from $q(\mathrm{z}_t|\mathrm{z}_{t-1}) = \mathcal{N}(\mathrm{z}_t; \sqrt{1-\beta_t}\mathrm{z}_{t-1}, \beta_t\mathbf{I})$, where β_t control the noise added at time step t. When T is large enough, a real-world sample will gradually and ultimately diffuse to a standard Gaussian noise distribution. The forward process is typically characterized by its simplicity and tractability, as it transforms the data into a noisy distribution that is easy to sample from a Gaussian distribution:

$$q(\mathrm{z}_t|\mathrm{z}_0) = \mathcal{N}(\mathrm{z}_t; \sqrt{\bar{\alpha}_t}\mathrm{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

For the reverse process, the diffusion model uses a learned parameterized denoising distribution $\mathrm{z}_{t-1} \sim p_\theta(\mathrm{z}_{t-1}|\mathrm{z}_t)$ to recover samples from noise gradually. The denoising distribution is parameterized by θ to fit the posterior distribution $q(\mathrm{z}_{t-1}|\mathrm{z}_t, \mathrm{z}_0)$ of the forward process. $q(\mathrm{z}_{t-1}|\mathrm{z}_t, \mathrm{z}_0)$ can be derived as:

$$q(\mathrm{z}_{t-1}|\mathrm{z}_t, \mathrm{z}_0) = \mathcal{N}(\mathrm{z}_{t-1}; \tilde{\mu}(\mathrm{z}_0, \mathrm{z}_t), \tilde{\beta}_t\mathbf{I}), \tag{2}$$

where $\tilde{\mu}(\mathrm{z}_0, \mathrm{z}_t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathrm{z}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathrm{z}_t$, and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. With learned denoising distribution p_θ, a synthetic real-world sample z_0 can be generated from pure random noise z_T step-by-step.

The primary objective of the loss function in diffusion models is to guide the reverse process in accurately recovering the original data from its noised version. One common formulation of the loss function in diffusion models is based on optimizing the Variational Lower Bound (VLB). A specific instantiation of the VLB optimization is the mean squared error (MSE) loss between the actual noise added to the data in the forward process and the noise predicted by the model in the reverse process. This MSE loss can be mathematically represented as follows:

$$\mathcal{L}_{\text{simple}}(\mathrm{z}_0) = \sum_{t=1}^{T} \mathbb{E}_{q(\mathrm{z}_t|\mathrm{z}_0)} ||\mu_\theta(\mathrm{z}_t, t) - \tilde{\mu}(\mathrm{z}_t, \mathrm{z}_0)||^2. \tag{3}$$

Sampling strategy DDIMs [21], an implicit generative model trained with denoising auto-encoding score matching objectives. DDIM can generate high-quality samples much more efficiently than existing DDPMs [17] and NCSNs [49], with the ability to perform meaningful interpolations from the latent space. In generative process, we can generate a sample z_{t-1} from a sample z_t by [21]:

$$\mathrm{z}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathrm{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\mathrm{z}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(\mathrm{z}_t) + \sigma_t\epsilon_t. \tag{4}$$

When $\sigma_t = \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$ for all t, the forward process becomes Markovian, and the generative process becomes a DDPM. Therefore, we downsample the generative process from 1000 to 200, which greatly speeds up our report generation algorithm without hurting sample quality.

3.2 Radiology report generation based on diffusion model

The task of radiology report generation necessitates that the model crafts an accurate diagnostic report $\mathcal{R} = \{w^1, w^2, ..., w^N\}$ of length N based on the provided radiographic images I^{radio}. Contrary to most previous approaches that employ autoregressive frameworks, which suffer from unsatisfactory inference speed and sequential error accumulation [13], our work utilizes diffusion models for this task. This process begins with noise that has been randomly sampled. Conditioned on the extracted features of radiographic images, it incrementally converts these random noises into the words that comprise radiology reports through a denoising process.

To tame diffusion models for a better text generator, we firstly extend continuous diffusion models to text with a discrete categorical nature, following the Diffusion-LM [18]. For a given radiology report, we map the discrete words to vectors $\{v^1, v^2, ..., v^N\}$ in continuous semantic space using an embedding function $v^i = g_\phi(w^i)$, where $v^i \in \mathbb{R}^d$ and ϕ represents the parameters of the embedding function. The inverse process involves mapping vectors v back to tokens w by selecting the most probable word at each position donated as $\tilde{p}(w|\mathrm{z}_0)$.

To enhance convergence of the generated z_0 towards a single word, we adopt the strategy from Diffusion-LM [18] by shifting from noise prediction to z_0 prediction. We employ a

modified loss function for the joint optimization of the diffusion model and word embedding function parameters to facilitate end-to-end training:

$$\begin{aligned}\mathcal{L}_{\text{simple}}(w) =& \mathbb{E}_{q_\phi(\mathrm{z}_0, h_{\hat{I}}, w)}[\sum_{t=2}^{T} \mathbb{E}_{q(\mathrm{z}_t|\mathrm{z}_0)} ||\mathrm{z}_\theta^0(\mathrm{z}_t, h_{\hat{I}}, t) - \mathrm{z}_0||^2 + \\ & ||\tilde{\mu}(\mathrm{z}_T, \mathrm{z}_0)||^2 + ||\mathrm{z}_\theta^0(\mathrm{z}_1, h_{\hat{I}}, 1) - g_\phi(w)||^2 - \log \tilde{p}(w|\mathrm{z}_0)],\end{aligned} \tag{5}$$

where z_θ is the denoising transformer, $h_{\hat{I}}$ is the feature of radiology image.

3.3 Graph-guided image feature extractor

Unlike conventional images, radiology images exhibit a greater similarity in appearance due to the imaging methods and human tissues themselves, and the areas of pathology are small and difficult to discern. Therefore, we employ prior medical knowledge to improve the extraction of features from radiology images.

Radiology image encoder Given a radiology image $I \in \mathbb{R}^{H\times W\times C}$, we firstly use the pre-trained medical visual encoder of MedClip [50] to extract the image features, where H, W, and C are the height, width, and number of channels of the image. Specifically, the image is initially split into patches with the shape of $P \times P$, which are then flattened into sequences of length S, where $\sqrt{S} = H/P = W/P$. After that, L^{Img} transformer blocks are employed to extract features from the image, where one encoding layer can be articulated as:

$$\bar{h}_I^l = \text{LN}(\text{MHA}(h_I^l) + h_I^l), \tag{6}$$

$$h_I^{l+1} = \text{LN}(\text{FFN}(\bar{h}_I^l) + \bar{h}_I^l), \tag{7}$$

where $h_I^l \in \mathbb{R}^{S\times d}$ is the input of the l-th layer. FFN and LN denote the Feed Forward Network [51] and Layer Normalization operation [52]. MHA [51] is multi-head attention. The image features outputted by the image encoder are denoted as h_I.

Medical graph encoder The chest knowledge graph G_{med} proposed in [1] has been widely integrated with radiology report generation systems to enhance the understanding of the relationships among pathologies. G_{med} consists of 27 entities and a root node referring to the global feature and an adjacency matrix $A_{\text{adj}} = \{e_{ij}\}$ to represent the edges. Each node is a disease keyword and we set e_{ij} to 1 when source node n_i connects target node n_j. Nodes linked to the same organ or tissue are connected and the root. Then we adopt pre-trained BERT [53] to initialize each entity and represent them as $h_G^1 \in \mathbb{R}^{27\times d}$. We employ L^{grp} transformer blocks to encode the features of the knowledge graph:

$$\bar{h}_G^l = \text{LN}(\text{RSA}(h_G^l, A_{\text{adj}}) + h_G^l), \tag{8}$$

$$h_G^{l+1} = \text{LN}(\text{FFN}(\bar{h}_G^l) + \bar{h}_G^l). \tag{9}$$

The relational self-attention (RSA) module is employed to integrate the structural graph information into the model. Specifically, the adjacency matrix, denoted as A_{adj}, functions as a visibility mask within the standard self-attention mechanism. This ensures that a node influences only its directly connected nodes, thereby reinforcing their interconnections.

Graph attention The graph attention aims to integrate knowledge from G_{med} with visual features. Following Liu et al. [54], we utilize L^{crs} cross-attention blocks to achieve this goal. The whole process can be written as follows:

$$\bar{h}_{\hat{I}}^l = \text{LN}(\text{CA}(h_{\hat{I}}^l, h_G) + h_{\hat{I}}^l), \tag{10}$$

$$h_{\hat{I}}^{l+1} = \mathrm{LN}(\mathrm{FFN}(\bar{h}_{\hat{I}}^{l}) + \bar{h}_{\hat{I}}^{l}). \tag{11}$$

In each cross attention head, *Query* comes from visual features $h_{\hat{I}}^{l}$, where $h_{\hat{I}}^{1} = h_I$. *Key* and *Value* come from the learned graph representations h_G.

3.4 Auxiliary lesion classification loss

Accurately categorizing pathology based on radiology image features is paramount in radiology report generation. To enhance the modeling of pathological features within radiology images, we introduce an auxiliary lesion classification loss that leverages constructed category information. Specifically, we utilize CheXbert [55] to generate a pseudo label for each image-text pair according to the radiology report $\mathcal{R}$ which can then be formulated as:

$$\{y_1, y_2, \cdots, y_i, \cdots, y_{N^{\mathrm{LC}}}\} = f_{\mathrm{lab}}(\mathcal{R}), \tag{12}$$

where the result is an one-hot vector and $y_i \in \{0, 1\}$ is the prediction result for i-th category. Note that the value of one indicates that category's existence, $N^{\mathrm{LC}} = 14$ is the number of categories, and f_{lab} denotes the automatic radiology report labeler.

Then we apply mean pooling to the output of the graph-guided image feature extractor $h_{\hat{I}}$, yielding $\tilde{h}_{\hat{I}} \in \mathbb{R}^d$. Subsequently, a linear classification head F_{cls} is utilized to conduct multi-label binary classification, after which a sigmoid function is applied to obtain the predicted probabilities, denoted as

$$\hat{y} = \mathrm{Sigmoid}(F_{\mathrm{cls}}(\tilde{h}_{\hat{I}})). \tag{13}$$

The auxiliary lesion classification is computed by the binary cross entropy (BCE):

$$\mathcal{L}_{\mathrm{ALC}} = \frac{1}{N^{\mathrm{LC}}} \sum_{i=1}^{N^{\mathrm{LC}}} \mathrm{BCE}(\hat{y}_i, y_i). \tag{14}$$

Finally, the model is jointly trained using both $\mathcal{L}_{\mathrm{ALC}}$ and $\mathcal{L}_{\mathrm{simple}}(w)$, culminating in a final loss function denoted as:

$$\mathcal{L}_{\mathrm{final}} = \mathcal{L}_{\mathrm{simple}}(w) + \mathcal{L}_{\mathrm{ALC}}. \tag{15}$$

4 Experiments

4.1 Datasets and evaluation metrics

Datasets We conduct experiments on two widely-used benchmarks for radiology report generation: IU-Xray [22] from Indiana University and MIMIC-CXR [23] from Beth Israel Deaconess Medical Center. The former dataset is a relatively small dataset with 7470 chest X-ray images and 3955 corresponding reports, the latter one is the largest public radiology dataset with 473057 chest X-ray images and 206563 reports. Following the experiment settings from previous studies [2, 5, 54], we exclusively focus on generating the findings section, while omitting samples lacking this section within both datasets. Specifically, for the IU-Xray dataset, we use the same split as stated in [5] and for MIMIC-CXR we adopt its official split. Table 1 shows the statistics of bath datasets in terms of the numbers of images, reports, patients, and the average length of reports with respect to training/validation/test set.
Evaluation metrics Following previous works [5, 56, 57], we employ the widely-used natural language generation (NLG) metrics to evaluate the quality of the generated radiology reports.

Table 1 The statistics of two benchmark datasets w.r.t their training, validation test sets, including the numbers of images, reports, and patients, and the averaged word-based length (AVG. LEN.) of reports

DATASET	IU-Xray			MIMIC-CXR		
	TRAIN	VAL	TEST	TRAIN	VAL	TEST
IMAGE	5.2K	0.7K	1.5K	369.0K	3.0K	5.2K
REPORT	2.8K	0.4K	0.8K	222.8K	1.8K	3.3K
PATIENT	2.8K	0.4K	0.8K	64.6K	0.5K	0.3K
AVG. LEN.	37.6	36.8	33.6	53.0	53.1	66.4

We adopt the standard evaluation protocol to calculate the captioning metrics: BLEU [58], ROUGE-L [59] and CIDEr [60].

4.2 Implementation details

We utilize a BART [61] structure as the denoising decoder, initialized randomly, and incorporated a modification where the causal attention mechanism is replaced with conventional self-attention. In the graph-guided image feature extractor, we use the pretrained MedClip [50] as the radiology image encoder. We employ the pretrained BERT [53] to encode the nodes within medical graphs. For training our method, We set the diffusion embedding dimension to 128 and the maximum diffusion step T to 1000. We use the $sqrt$ schedule from DiffusionLM [18] to initialize the adaptive time schedule with a learning rate of 10e-4 with 10000 warm-up step and a linearly-decreasing schedule. The proposed adaptive noise schedule is updated every 10000 training steps for IU-Xray and 20000 for MIMIC-CXR. During the generative process, we utilize 200 inference steps and explore maximum Bayes risk (MBR) decoding [62] to select the best sample. We trained our model on 2 NVIDIA 3090 Ti GPUs and evaluated the performance on 1 NVIDIA 3090 Ti GPU.

4.3 Main results

We compare DMR^{2}G with a wide range of existing state-of-the-art autoregressive methods on two benchmarks. R2Gen [5] and R2GenCMN [63] have been widely used a baseline MRG model recently. PPKED [54] and MET [57] are proposed to integrate medical knowledge with typical MRG backbones. XProNet [56] utilize cross-modal prototypes to record the information. Moreover, we also compare the semi-autoregressive (Semi-AR) and NAR methods. SATIC [64] keeps the autoregressive property in global but generates words parallelly in local. CMAL [65] utilize a multi-agent reinforcement learning system to cooperatively maximize a sentence-level reward. In addition, we compare the inference speed with some methods that have publicly released code.

Given that the prevailing paradigm adopts an autoregressive framework, we categorize autoregressive models into a single group for comparative analysis, while Semi-AR and NAR models are categorized into one group for comparison. As shown in Table 2, the proposed DMR^{2}G demonstrates comparable report quality to that of previous state-of-the-art autoregressive models on both datasets but with a noticeable speedup. In detail, it outperforms classic image captioning methods such as Att2In [66] in terms of both performance and inference speed. DMR^{2}G also achieves considerable performances over autoregressive models such as R2Gen [5] while maintaining an approximately 3.64× speed advantage on the IU-Xray dataset and 7.37× speed advantage on the MIMIC-CXR dataset. Compared to the non-autoregressive models, our method strikes a better balance between performance and speed. In summary, our proposed method demonstrates a superior balance between

Table 2 Performance of our proposed DMR^{2}G and other state-of-the-art methods on the IU-Xray and MIMIC-CXR datasets

		IU-Xray [22]							
Methods	Pattern	B-1	B-2	B-3	B-4	R-L	C	Latency	Speedup
Att2In [66]	AR	0.248	0.134	0.116	0.091	0.309	0.215	256 ms	0.55×
R2Gen [5]		0.470	0.304	0.219	0.165	0.371	0.398	142 ms	1.00×
PPKED [54]		**0.483**	0.315	0.224	0.168	0.376	0.351	–	–
R2GenCMN [63]		0.474	0.302	0.220	0.168	0.370	–	64 ms	2.22×
XProNet [56]		0.463	0.301	0.210	0.156	0.359	–	**61 ms**	**2.33** ×
MET [57]		**0.483**	**0.322**	**0.228**	**0.172**	**0.380**	**0.435**	–	–
SATIC [64]	Semi-AR	0.424	0.286	0.169	0.145	0.337	0.349	137 ms	1.04×
DSA-Transformer [67]		**0.466**	**0.303**	**0.219**	**0.166**	**0.372**	**0.391**	117 ms	1.21×
CMAL [65]	NAR	0.232	0.116	0.101	0.083	0.286	0.198	96 ms	1.48×
DMR^{2}G (Ours)		0.465	0.293	0.202	0.146	0.360	0.381	**39 ms**	**3.64**×
		MIMIC-CXR [23]							
Methods	Pattern	B-1	B-2	B-3	B-4	R-L	C	Latency	Speedup
Att2In [66]	AR	0.314	0.198	0.133	0.095	0.264	0.106	–	–
R2Gen [5]		0.353	0.218	0.145	0.103	0.277	0.253	435 ms	1.00×
PPKED [54]		0.360	0.224	0.149	0.106	0.284	0.237	–	–
R2GenCMN [63]		0.353	0.218	0.148	0.106	0.278	–	131 ms	3.32×
XProNet [56]		0.344	0.215	0.146	0.105	0.279	–	**112 ms**	**3.88**×
MET [57]		**0.386**	**0.250**	**0.169**	**0.124**	**0.291**	**0.362**	–	–
SATIC [64]	Semi-AR	0.364	0.208	0.133	0.089	0.266	0.248	401 ms	1.08×
DSA-Transformer [67]		–	–	–	–	–	–	–	–
DMR^{2}G (Ours)	NAR	**0.373**	**0.217**	**0.136**	**0.099**	**0.267**	**0.262**	**59 ms**	**7.37**×

A higher value denotes better performance in all columns. The best score is highlighted in **bold**

performance and inference speed compared to other autoregressive or non-autoregressive approaches.

4.4 Ablation study

To fully investigate the contribution of our proposed graph-guided image feature extractor, auxiliary lesion classification loss, and different inference steps, we perform ablation experiments on both datasets. Our base model only keeps the transformer structure and employs the diffusion model with 200 inference steps. The experimental results are shown in Table 3.

Effectiveness of graph-guided image feature extractor For the graph-guided image feature extractor, we remove the graph-guided image feature extractor (GIFE) module and only retain the radiology image encoder to extract image features. The medical graph encoder utilizes prior knowledge to aid in the identification of abnormal regions within the image. As shown in Table 3, when removing the GIFE module, our model has a significant performance drop on both IU-Xray and MIMIC-CXR datasets, especially in CIDEr score. The results demonstrate that the graph can enhance the understanding of the relationships among pathologies and facilitate a more precise extraction of image features.

Table 3 The experimental results of ablation studies on the IU-Xray and MIMIC-CXR datasets

IU-Xray	B-1	B-2	B-3	B-4	R-L	C
w/o ALCL	0.441	0.279	0.192	0.139	0.356	0.333
w/o GIFE	0.454	0.282	0.188	0.129	0.348	0.246
DDIM 50 Steps	0.416	0.255	0.169	0.119	0.349	0.295
DMR^{2}G	**0.465**	**0.293**	**0.202**	**0.146**	**0.360**	**0.381**
MIMIC-CXR	B-1	B-2	B-3	B-4	R-L	C
w/o ALCL	0.353	0.201	0.125	0.091	0.248	0.249
w/o GIFE	0.362	**0.226**	0.132	0.095	0.256	0.211
DDIM 50 Steps	0.342	0.199	0.123	0.082	0.250	0.245
DMR^{2}G	**0.373**	0.217	**0.136**	**0.099**	**0.267**	**0.262**

The best values are highlighted in **bold**

Effectiveness of auxiliary lesion classification loss We remove the auxiliary lesion classification loss (ALCL) that augments the modeling of pathological features within radiology images. As a supervisory signal, the loss employs pseudo labels to align image abnormal region features and textual disease keywords. In the absence of this particular module, as shown in Table 3, the BLEU-1 scores decline from 0.465 to 0.441 and from 0.373 to 0.353 in both datasets, respectively. The results demonstrate the effectiveness of the auxiliary lesion classification loss.

Effectiveness of less inference steps During the generative process, we reduce the inference steps from 200 to 50. The reduction in inference steps inherently leads to expedited inference speed. Conversely, the diminished inference steps engender challenges in ensuring the predicted text representation can be accurately mapped back to discrete words. As shown in Table 3, the performance metrics demonstrate a comprehensive decline when employing 50 inference steps. Consequently, we opt to utilize 200 inference steps to achieve a more favorable equilibrium between performance and speed.

4.5 Case study

To further investigate the effectiveness of our method, we perform qualitative analysis on MIMIC-CXR [23] with their ground truth, XProNet [56], and generate reports from different models. As we can see in Fig. 3, DMR^{2}G is able to generate descriptions aligned with those written by radiologists with similar contents. For the autoregressive framework, XProNet

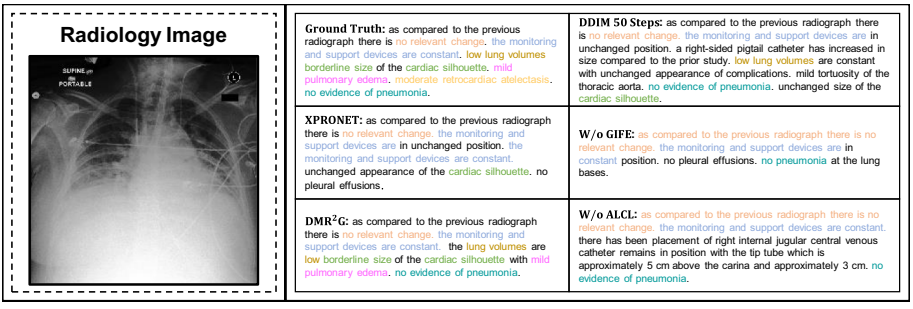

Fig. 3 Illustrations of the reports generated by different models for one sample from MIMIC-CXR [23]. For better visualization, different colors indicate different symptoms

suffers from sequential error accumulation [13] leading to generating the duplicate sentences, as highlighted in the blue-marked sentences of the result. In contrast, our method does not suffer from this issue.

In addition, this case serves as a qualitative analysis of the ablation experiments. For example, following the exclusion of the medical graph, DMR^{2}G exhibits a diminished interrelation among medical entities, thereby leading to instances where associated entities fail to co-occur concurrently. The exclusion of the abnormal classification loss within DMR^{2}G attenuates the concordance between abnormal regions in images and their corresponding pathological entities, potentially introducing bias into the process of report generation. When we use fewer sampling steps for the model to generate reports, the fewer sampling steps lead to a decrease in the quality of the generation.

5 Conclusion

In this paper, we present DMR^{2}G, a novel framework based on diffusion models for radiology report generation. We propose the graph-guided image feature extractor that utilizes the prior knowledge embedded within the pre-construct graph to aid in the identification of abnormal regions within the image. Additionally, we also propose the auxiliary lesion classification loss to align image abnormal region features and textual disease keyword representations by employing pseudo labels. Through the utilization of the accelerated sampling strategy inherent to diffusion models, our method effectively mitigates the constraints inherent in conventional autoregressive models. Experimental results on the IU-Xray and MIMIC-CXR datasets demonstrate that our proposed method achieves a better trade-off between performance and inference speed than other approaches.

Author Contributions The authors confirm contribution to the paper as follows: study conception, design, analysis and interpretation of results: Huan Ouyang, Zheng Chang; draft manuscript preparation: Huan Ouyang, Zheng Chang, Binghao Tang; Direction: Si Li. All authors reviewed the results and approved the final version of the manuscript.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics for Obtaining the Data In this paper, all the conditions specified for the use of the open datasets taken as a source for the generative process are satisfied.

References

1. Zhang Y, Wang X, Xu Z, Yu Q, Yuille A, Xu D (2020) When radiology report generation meets knowledge graph. In: Proceedings of the AAAI conference on artificial intelligence 34:12910–12917
2. Li Y, Liang X, Hu Z, Xing EP (2018) Hybrid retrieval-generation reinforced agent for medical image report generation. Adv Neural Inf Process Syst 31
3. Wang X, Peng Y, Lu L, Lu Z, Summers RM (2018) Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9049–9058

4. Jing B, Wang Z, Xing E (2020) Show, describe and conclude: on exploiting the structure information of chest x-ray reports. arXiv:2004.12274
5. Chen Z, Song Y, Chang T-H, Wan X (2020) Generating radiology reports via memory-driven transformer. arXiv:2010.16056
6. Liu F, Yin C, Wu X, Ge S, Zou Y, Zhang P, Sun X (2021) Contrastive attention for automatic chest x-ray report generation. arXiv:2106.06965
7. Jing B, Xie P, Xing E (2017) On the automatic generation of medical imaging reports. arXiv:1711.08195
8. Liang X, Hu Z, Zhang H, Gan C, Xing EP (2017) Recurrent topic-transition GAN for visual paragraph generation
9. Li M, Liu R, Wang F, Chang X, Liang X (2023) Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web 26(1):253–270
10. Chen Y-J, Shen W-H, Chung H-W, Chiu C-H, Juan D-C, Ho T-Y, Cheng C-T, Li M-L, Ho T-Y (2022) Representative image feature extraction via contrastive learning pretraining for chest x-ray report generation. arXiv:2209.01604
11. Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P (2021) Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine learning for health, PMLR, pp 209–219
12. Liu F, Yin C, Wu X, Ge S, Zou Y, Zhang P, Sun X (2021) Contrastive attention for automatic chest x-ray report generation. arXiv:2106.06965
13. Gao J, Meng X, Wang S, Li X, Wang S, Ma S, Gao W (2019) Masked non-autoregressive image captioning
14. Gu J, Bradbury J, Xiong C, Li VO, Socher R (2017) Non-autoregressive neural machine translation. arXiv:1711.02281
15. Qin L, Welleck S, Khashabi D, Choi Y (2022) COLD decoding: energy-based constrained text generation with Langevin dynamics
16. Ren Y, Liu J, Tan X, Zhao Z, Zhao S, Liu T-Y (2020) A study of non-autoregressive model for sequence generation. arXiv:2004.10454
17. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851
18. Li X, Thickstun J, Gulrajani I, Liang PS, Hashimoto TB (2022) Diffusion-lm improves controllable text generation. Adv Neural Inf Process Syst 35:4328–4343
19. Gong S, Li M, Feng J, Wu Z, Kong L (2022) Diffuseq: Sequence to sequence text generation with diffusion models. arXiv:2210.08933
20. Liu G, Li Y, Fei Z, Fu H, Luo X, Guo Y (2023) Prefix-diffusion: a lightweight diffusion model for diverse image captioning
21. Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. arXiv:2010.02502
22. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ (2016) Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inf Assoc 23(2):304–310
23. Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng C-y, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019) Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv:1901.07042
24. Chen X, Fang H, Lin T-Y, Vedantam R, Gupta S, Dollar P, Zitnick CL (2015) Microsoft COCO captions: data collection and evaluation server
25. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator
26. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering
27. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions
28. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning
29. Huang L, Wang W, Chen J, Wei X-Y (2019) Attention on attention for image captioning
30. Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: adaptive attention via a visual sentinel for image captioning
31. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2016) Show neural image caption generation with visual attention, attend and tell
32. Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning
33. Yao T, Pan Y, Li Y, Mei T (2019) Hierarchy parsing for image captioning
34. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2017) Semantic compositional networks for visual captioning
35. Wu Q, Shen C, Liu L, Dick A, Hengel A (2016) What value do explicit high level concepts have in vision to language problems?

36. Yao T, Pan Y, Li Y, Qiu Z, Mei T (2016) Boosting image captioning with attributes
37. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention
38. Yang X, Tang K, Zhang H, Cai J (2018) Auto-encoding scene graphs for image captioning
39. Wang Z, Liu L, Wang L, Zhou L (2023) R2GenGPT: radiology report generation with frozen LLMs. arXiv:2309.09812
40. Liu C, Tian Y, Chen W, Song Y, Zhang Y (2024) Bootstrapping large language models for radiology report generation. In: Wooldridge MJ, Dy JG, Natarajan S (eds.) AAAI, pp 18635–18643
41. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
42. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv:1312.6114
43. Ho J, Saharia C, Chan W, Fleet DJ, Norouzi M, Salimans T (2022) Cascaded diffusion models for high fidelity image generation. J Mach Learn Res 23(1):2249–2281
44. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst 34:8780–8794
45. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2021) Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv:2112.10741
46. Hoogeboom E, Gritsenko AA, Bastings J, Poole B, Berg Rvd, Salimans T (2021) Autoregressive diffusion models. arXiv:2110.02037
47. Chen T, Zhang R, Hinton G (2022) Analog bits: Generating discrete data using diffusion models with self-conditioning. arXiv:2208.04202
48. He Y, Cai Z, Gan X, Chang B (2023) DiffCap: exploring continuous diffusion on image captioning
49. Song Y, Ermon S (2020) Generative modeling by estimating gradients of the data distribution
50. Wang Z, Wu Z, Agarwal D, Sun J (2022) Medclip: contrastive learning from unpaired medical images and text. arXiv:2210.10163
51. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
52. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization
53. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding
54. Liu F, Wu X, Ge S, Fan W, Zou Y (2021) Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13753–13762
55. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP (2020) Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv:2004.09167
56. Wang J, Bhalerao A, He Y (2022) Cross-modal prototype driven network for radiology report generation
57. Wang Z, Liu L, Wang L, Zhou L (2023) METransformer: radiology report generation by transformer with multiple learnable expert tokens
58. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318
59. Lin C-Y (2004) Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
60. Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4566–4575
61. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461
62. Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 388–395
63. Chen Z, Shen Y, Song Y, Wan X (2022) Cross-modal memory networks for radiology report generation
64. Zhou Y, Zhang Y, Hu Z, Wang M (2021) Semi-autoregressive transformer for image captioning
65. Guo L, Liu J, Zhu X, He X, Jiang J, Lu H (2020) Non-autoregressive image captioning with counterfactuals-critical multi-agent learning
66. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
67. Tang Y, Wang D, Zhang L, Yuan Y (2024) An efficient but effective writer: diffusion-based semi-autoregressive transformer for automated radiology report generation. Biomed Signal Process Control 88:105651