Diversity-Augmented Diffusion Network with LLM Assistance for Radiology Report Generation

Jieting Long
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
jlon5443@uni.sydney.edu.au

Zhiyuan Li
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
zhli0736@uni.sydney.edu.au

Jianan Fan
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
jianan.fan@sydney.edu.au

Zhuonan Liang
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
zhuonan.liang@sydney.edu.au

Ao Ma
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
aoma0081@uni.sydney.edu.au

Henning Müller*†
The University of Applied Sciences
Western Switzerland (HES-SO)
Institute of Informatics
Sierre, Switzerland
henning.mueller@hevs.ch

Weidong Cai
The University of Sydney
School of Computer Science
Sydney, NSW, Australia
tom.cai@sydney.edu.au

Abstract

Radiology report generation (RRG) is a demanding yet challenging task that involves producing multi-sentence diagnostic narratives, requiring long-form text with high diversity while addressing inherent data bias. Sentence-level diversity is therefore crucial for capturing varying diagnostic details across multiple regions of interest (ROIs) within a single report, yet it remains underexplored in the field. In this paper, we propose DADNET, a novel diffusion-based framework that leverages the inherent ability of diffusion models to generate diverse text. We make the first attempt to integrate large language models (LLMs) to bridge the inherent training-inference gap in diffusion models. Specifically, LLMs are used to draft a preliminary report, which is subsequently incorporated into the diffusion process to enhance report diversity. Additionally, we introduce a bias equalization technique using domain-specific priors to mitigate data distribution biases, improving the quality and reliability of generated reports under various scenarios. Experimental results demonstrate that DADNET outperforms existing approaches under the same non-autoregressive (NAR) mechanism and sets a new benchmark for diversity in RRG. This work underscores the importance of diversity in RRG and establishes DADNET as a pioneering framework for addressing this challenge with NAR methods.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Natural language generation**.

Keywords

Radiology Report Generation, Large Language Model, Diffusion Model

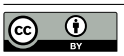
ACM Reference Format:
Jieting Long, Zhiyuan Li, Jianan Fan, Zhuonan Liang, Ao Ma, Henning Müller, and Weidong Cai. 2025. Diversity-Augmented Diffusion Network with LLM Assistance for Radiology Report Generation. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3701716.3717555

1 Introduction

The growing disparity between the volume of radiological imaging data and the availability of professional readers has resulted in an unsustainable escalation in radiologists' workloads. This mounting pressure to manage an increasing number of complex cases within tight time constraints often compromises the quality of diagnostic reports. These concerns underscore the necessity for automated radiology report generation (RRG) solutions, which promise to alleviate radiologists' workload, reduce diagnostic errors, and streamline clinical workflows.

Recent deep learning-based RRG works [3, 10, 17, 43] are typically built on the advancements in natural image captioning methods, adopting an encoder-decoder autoregressive (AR) framework, integrating modifications specifically tailored to the distinct challenges posed by RRG tasks. Specifically, these improvements are designed to address two primary challenges that set RRG apart from conventional image captioning.

*Also with Department of Radiology and Medical Informatics, The University of Geneva, Switzerland.
†Also with The Sense Innovation and Research Center, Lausanne & Sion, Switzerland.



WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717555

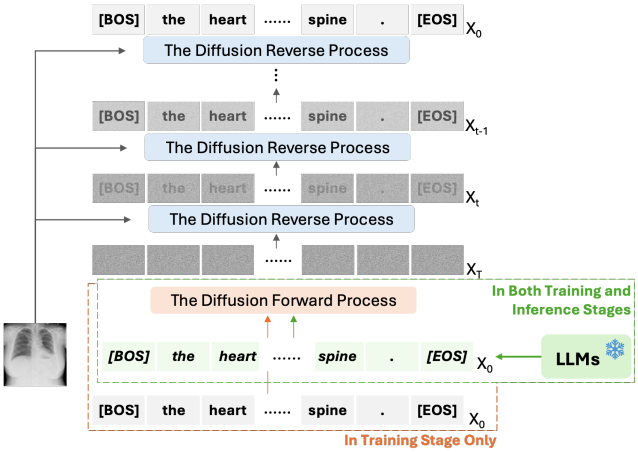


Figure 1: An illustration describing the workings of the conventional diffusion-based RRG framework and our proposed approach, highlighting their unified structures and the key difference in-between. The orange elements represent the traditional approach, while the green components showcase our main contribution with LLM. The reverse process, as shown, is conceptually similar across both frameworks. Best viewed in colour for clarity.

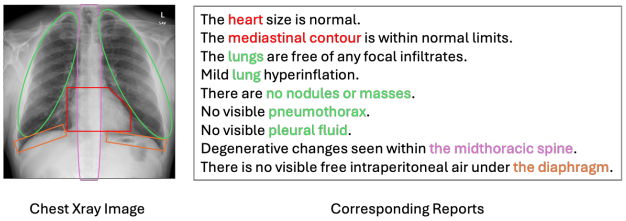


Figure 2: Sentence-Level Diversity Characteristics in RRG Tasks. The highlighted phrases correspond to specific regions of interest (ROIs) in the radiograph, marked in matching colors for clarity. Best viewed in color.

The first challenge lies in producing long-text descriptions with high diversity, a key feature that distinguishes RRG apart from natural image captioning tasks. While image captions are typically concise and focus on summarizing a single object or scene, radiology reports are significantly longer, comprising multiple sentences that describe various organs or regions of interest, as demonstrated in Figure 2. Each sentence in a report may describe distinct findings, with diagnoses for each region being either interrelated or entirely independent. The inherent diversity and complexity require models to capture long-range dependencies across sentences while maintaining contextual accuracy and variability in their outputs. To tackle these challenges, several approaches have been developed. Notably, the successful introduction of memory mechanisms in models like R2Gen [4], R2GenCMN [3] and their invariants [10, 28, 37] has demonstrated the importance of enabling models to effectively handle long-form content for RRG tasks. However, despite these advancements, limited attention has been given to discussing the degree of diversity in generated reports, exposing a critical gap in current research.

The second challenge stems from imbalanced information within both visual and textual modalities. The data bias occurs as inter-sample imbalance, where normal samples dominate datasets despite abnormal ones being diagnostically crucial, and also intra-sample imbalance, where abnormal findings are confined to small portions of the data for both modalities (e.g., negative descriptions usually consisting of brief and specific phrases featuring medical terminology), leaving irrelevant information to take up the majority of the space. The issue becomes particularly intricate for textual data. Radiology reports often adhere to standardized templates to ensure clarity and practicality for clinical use. While template words may appear redundant and less informative for learning abnormal descriptions, they are essential for generating standardized and clinically usable reports. On the other hand, specialized medical terminologies critical for diagnostic accuracy may be inadequately represented during training, resulting in limited model learning. This underscores the demand for a model capable of capturing a greater degree of diversity. To tackle these challenges, there has been some pioneering efforts, such as aligning features across modalities to emphasize diagnostically relevant regions [44] or refining features using external knowledge, including report templates or medical knowledge graphs, to guide the learning process [22]. Despite these advancements demonstrate their effectiveness within transformer-based framework, they lack sufficient investigation and exploration of alternatvie decoding paradigms, highlighting the need for further research and innovation.

In this paper, recognizing the significance of diversity in RRG and the insufficient attention it has received, we prioritize a more fundamental aspect to enhancing diversity by rethinking the underlying learning paradigm, rather than pursuing incremental advancements through auxiliary module additions. Specifically, we move away from widely explored transformer-based encoder-decoder architectures to a diffusion-based framework. This transition is driven by two primary motivations. First, diffusion-based methods inherently promote diversity in text generation by enabling bidirectional message passing and parallel token generation [9]. In contrast, AR models generate tokens sequentially in a unidirectional manner, which often results in error accumulation and reduced diversity [31]. Second, diffusion-based methods offer greater efficiency, as their parallel processing allows for faster generation and more effective handling of longer texts compared to the sequential nature of autoregressive approaches.

A major hurdle in applying diffusion-based techniques lies in the noise inconsistency between training and inference. During training, noise is added to ground-truth reports, retaining some underlying information and causing potential leakage [45], whereas inference starts with pure Gaussian noise, leading to performance gaps. To address this, we propose DADNET, a non-autoregressive framework leveraging large language models (LLMs). A lightly fine-tuned LLM acts as a "radiologist intern", producing a rough report draft used as a more informative starting point during inference and randomly as input during training. As illustrated in Figure 1, our proposed approach incorporates the LLM-assisted module, shown

in green, to effectively mitigate the noise gap, improve consistency, and boost overall performance. In addition, to tackle data bias, we integrate prior knowledge prompts into the report generation process, ensuring the outputs are both coherent and diagnostically accurate.

In sum, our contributions can be summarized as follows:

- We are the first to propose an RRG framework that incorporates a large language model (LLM) into a diffusion-based non-autoregressive paradigm;
- We leverage LLMs to generate a preliminary report draft, which helps bridge the gap between training and inference while enhancing the diversity of the generated reports;
- We are the first to highlight and prioritize the importance of diversity evaluation in the context of RRG tasks;
- **DADNET** outperforms multiple baselines under the non-autoregressive (NAR) mechanism in standard natural language generation (NLG) metrics, and it delivers exceptional results on diversity metrics when compared to both autoregressive and non-autoregressive methods.

2 Related Work

Radiology Report Generation. In terms of the model architecture, most radiology report generation methods utilize either CNN- or Transformer-based visual extractors combined with autoregressive text generators (i.e., RNN- or Transformer-based models). For instance, Zhou et al. [48] proposed a multi-modality semantic attention module along with additional topic-level losses within a CNN-RNN architecture to improve the precision of the generated reports. While R2Gen [4] introduced the concept of memory into a Transformer-based encoder-decoder structure, becoming the first method to employ Transformer architectures in this domain (for both visual and textual modalities). It has since served as a foundational starting point for further advancements in radiology report generation. Recently, D^2-Net makes the first attempt to apply diffusion-based models to radiology report generation, marking an initial step in exploring this approach for the task.

Given the similarities to natural image captioning, which has been extensively studied, many existing RRG approaches build upon these methods with domain-specific adaptations. These approaches can generally be categorized based on their strategies: enhancing learning patterns or incorporating external domain-specific knowledge. For example, R2Gen [4] improves the textual generation process by leveraging relational memory to capture pattern information. It learns relational memory from previous generation steps and integrates this information into normalization layers in the decoder, thereby enhancing the quality of the generated content. Similarly, R2GenCMN [3] expands on this concept by utilizing memory to improve cross-modal alignment, ensuring the generated reports are both semantically coherent and highly aligned with the input images. On the other hand, COMG [10] incorporates prior knowledge, such as diagnostic priors and organ masks, at both the pixel and textual levels to aid feature learning directly from radiographs, simplifying the report generation process.

The review of existing works reveals notable limitations in methodological flexibility and untapped opportunities. From the perspective of the decoding paradigm, since the introduction of R2Gen [4], Transformer-based architectures have dominated radiology report generation. While effective, the AR mechanism faces challenges like unidirectional message passing, which causes error accumulation degrading the content quality. The recent success of diffusion-based non-autoregressive frameworks [19] underscores the potential of exploring diffusion models as a promising alternative. From another perspective, the success of existing methods in improving feature learning — whether through refining learning patterns [3, 4, 43, 48] or integrating external domain-specific knowledge [10, 33] — highlights the importance of tackling task-specific challenges in RRG. These advancements point to untapped opportunities for further investigation of challenges specific to the RRG task.

Large Language Models in Report Generation. Large language models (LLMs) have demonstrated significant advantages in performing logical reasoning and generating coherent and contextually relevant texts, thanks to their extensive knowledge base and sophisticated linguistic understanding [6, 24, 38, 39]. These capabilities position LLMs as promising solutions for tasks requiring advanced linguistic proficiency, such as radiology report generation (RRG). Consequently, several studies have explored the use of LLMs for RRG [11, 22, 36, 39, 40, 43]. More specifically, R2GenGPT [43], KARGEN [22] and RaDialog [36] all adopt Llama2-based LLM as the report generator, and RGRG [40] adopts GPT2-Medium. Despite their contributions, these approaches share notable limitations. Most rely on LLMs strictly as report generators within transformer-based frameworks, which inherently depend on autoregressive mechanisms for language modeling. This reliance not only perpetuates issues such as error accumulation in sequential text generation but also requires substantial computational resources for loading LLMs and fine-tuning them to enhance domain-specific performance. Moreover, these methods fail to explore alternative architectures or innovative applications of LLMs beyond direct text generation. This gap highlights an exciting opportunity to reimagine the integration of LLMs into RRG workflows. For example, LLMs have been effectively utilized to generate domain-specific knowledge to support model learning in related tasks, such as medical visual question answering [21, 23, 27], task-oriented dialogue modeling [29] and medical report summarization [32]. In our proposed method, we extend the role of LLMs beyond traditional report generation by employing them as external agents to assist a diffusion-based non-autoregressive (NAR) framework.

3 Methods

In this section, we introduce our framework for radiology report generation, named Diversity-Augmented Diffusion Network (DADNET), which is established upon D2Net [19] in a diffusion-based image-to-text generation paradigm [31]. Our method reflects diversity-augmentation by integrating a large language model (LLM) to bridge the noise gap between training and inference within the diffusion-based framework. In addition, it employs prior-knowledge prompts to tackle typical data bias issues in RRG, leveraging medical domain knowledge to enhance the model's focus on diagnostically relevant information. The overall structure of DADNET is illustrated in Figure 3.

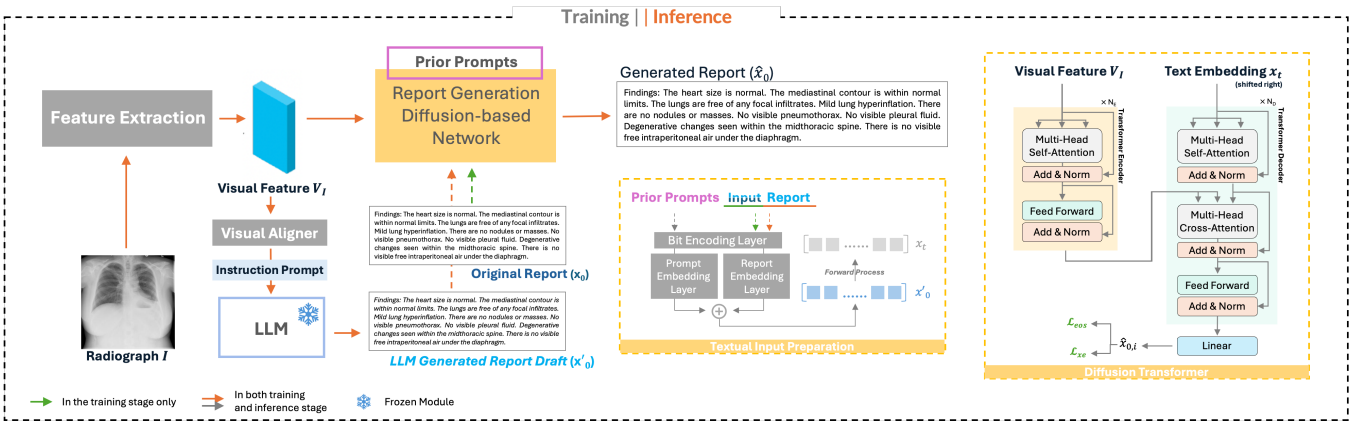


Figure 3: An overview of our proposed DADNET: (1) A diffusion model utilizing LLM-generated reports as input to address the noise gap in-between, and (2) Prior Knowledge prompts to incorporate domain-specific expertise into the diffusion process. Detailed explanations of each component are provided in Section 3.

3.1 Preliminary: Diffusion-based RRG

Input Preprocessing. Given the image modality input, a radiograph I, we leverage ResNet101, a CNN-based model, to extract visual features represented as $V_I = f_v(I)$. Here, f_v refers to the ResNet [12] model, and $V_I = v_1, v_2, \ldots, v_S$ represents the set of visual features extracted from the radiograph, where each feature $v_s \in \mathbb{R}^d$ has a dimensionality of d. While for text modality, an input report R, we leverage the Bit Encoding Layer [2] to encode it into $x_0 \in \mathbb{R}^{n \times L}$. Here, $n = \lceil \log_2 \mathcal{W} \rceil$ corresponds to the number of binary bits (i.e., $\{0,1\}^n$) required for the one-hot encoding of each token, where $\mathcal{W}$ denotes the vocabulary size of the corpus. The hyperparameter L represents the pre-defined maximum sequence length, which is set to ensure parallel computation by padding or truncating all report samples to a uniform length.

Diffusion Process. The encoded text input x_0 undergoes a diffusion process with forward and reverse steps. In the forward step, Gaussian noise is progressively added to x_0 for $t \in (0, T]$, resulting in the noisy input x_t, computed as:

$$x_t = \sqrt{\mathbf{sigmoid}(-\gamma(t'))}x_0 + \sqrt{\mathbf{sigmoid}(\gamma(t'))}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,1)$ represents Gaussian noise, and $t \sim \mathcal{U}(0,T)$ is a continuous time variable normalized to $t' = t/T$. In the reverse step, the noisy input x_t is denoised back to x_0, conditioned on visual features V_I, by training a diffusion transformer f_θ to minimize the following ℓ_2 loss:

$$\mathcal{L}_{bit} = \mathrm{E}_{t\sim\mathcal{U}(0,T),\epsilon\sim\mathcal{N}(0,1)}||f_\theta(x_t, \gamma(t'), V_I) - x_0||^2. \tag{2}$$

During inference, the reverse process begins with x_t initialized as pure Gaussian noise at $t = T$ and iteratively applies f_θ to generate latent states as t transitions to $t = 0$. The final x_0 is then obtained using the sampling strategy proposed in DDPM [13].

The diffusion transformer utilizes both the visual features V_I and the noise-injected text input x_t in an encoder-decoder structure. The decoder generates textual hidden states H_t conditioned on visual hidden states H_v produced by the encoder, to reconstruct a clean report $\hat{x_0}$. The process is operated as follows:

$$\begin{aligned}
H_v &= f_e(V_I) = \mathrm{FFd}(\mathrm{LNorm}(\mathrm{MultiAttn}(\mathbf{V_I}, \mathbf{V_I}, \mathbf{V_I})) + \mathbf{V_I}),\\
H_x &= f_d(H_v, x_t) = \mathrm{FFd}(\mathrm{LNorm}(\mathrm{MultiAttn}(h_{xt}, H_v, H_v)) + h_{xt}),\\
h_{xt} &= \mathrm{LNorm}(\mathrm{MultiAttn}(x_t, x_t, x_t) + x_t),\\
\mathrm{MultiAttn}(Q,K,V) &= \mathrm{CAT}(\mathrm{head}_0, \mathrm{head}_1, \ldots, \mathrm{head}_H)W_O,\\
\mathrm{head}_i &= \mathrm{Attn}(W_{qi}\mathbf{V_I}, W_{ki}\mathbf{V_I}, W_{vi}\mathbf{V_I}),\\
\mathrm{Attn}(q,k,v) &= \mathrm{softmax}\left(\frac{qk^T}{\sqrt{d}}\right)v,
\end{aligned} \tag{3}$$

where W_q, W_k, W_v are learnable parameters, MultiAttn denotes the multi-head attention module with a predefined number of heads, LNorm represents layer normalization, FFd is the feedforward layer. f_e and f_d refer to the encoder and decoder, while H_v and H_t are the visual and textual hidden states, respectively.

In addition to $\mathcal{L}_{bit}$ in Eq. 2, the diffusion process objective also incorporates two cross-entropy loss terms within each diffusion transformer block: $\mathcal{L}_{xe}$, which evaluates the accuracy of each output word, and $\mathcal{L}_{eos}$, which assesses whether each output word represents the end-of-sentence (i.e., its semantic meaningfulness).

3.2 LLM-Assisted Noise Initialization

The reverse step of diffusion models operates differently during training and inference, resulting in an implicit gap between the two stages. During training, the process begins with corrupted inputs generated from the original data through the forward diffusion process. However, in the inference stage, it typically starts with pure Gaussian noise. This mismatch in the initial noise introduces a gap, potentially leading to information leakage during training [7, 26, 45]. Recent studies [7, 8, 30, 45] further demonstrate that incorporating correlated initial noise during inference significantly enhances the quality of generated content in diffusion-based image and video generation tasks. Building on these insights, we propose leveraging

Table 1: NLG Analysis: Comparison with NAR method on IU-Xray

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
1-NN	0.232	0.116	0.051	0.018	-	0.201	-
HRGR	0.438	0298	0.208	0.151	-	0.322	-
D^2-Net†	0.434	0.283	0.197	0.129	0.177	0.326	0.389
DADNET	0.450	0.302	0.224	0.175	0.203	0.345	0.708

pretained LLMs to generate informed initial noisy reports tailored specifically for our diffusion-based report generation framework.

While RRG inherently involves a multi-modal approach, we opt to employ LLMs integrated with a visual aligner rather than leveraging pretrained visual-language models (VLMs) due to several key considerations. First, VLMs are typically trained on general-purpose datasets, which often lack the specialized anatomical and pathological nuances essential for understanding medical imaging. Second, while VLMs excel in visual-textual alignment [16], they generally underperform in text-based inferential reasoning [47]. In contrast, pretrained LLMs, particularly when fine-tuned with domain-specific instruction prompts, not only retain their strong inferential reasoning capabilities in the domain but also effectively accommodate multi-modal inputs tailored to the medical domain. In addition, the visual information processed by 'the official radiologist' (i.e., the diffusion-based generator) and 'the intern' (i.e., the LLM) should remain consistent to minimize discrepancies in image understanding between the two.

To generate such report, the extracted visual features V_I are also mapped into a higher-dimensional feature space compatible with the LLM to integrate visual information into the LLM. This mapping is achieved through a multi-layer perception, producing a sequence of visual tokens Z_v, referred to as the visual aligner:

$$Z_v = g_{vm}(V_I), \tag{4}$$

where $g_{vm}(\cdot)$ is the visual aligner.

For the LLM component, we adopt LLaMA2-7B [41] selected for its robust capabilities and effectiveness in handling complex language tasks. Drawing inspiration from [43], we design our instruction prompts X_{prompt} based on theirs as follows:

"`Patient: <Img>`Z_v`</Img>, Generate a comprehensive and detailed diagnosis report for this chest x-ray image. \n Radiologist:` X_r $<$ `/s` $>$."

Here, X_r denotes the corresponding report, with all prompt text tokenized using LLaMA's tokenizer for optimal processing.

Given the visual features V_I and instruction prompts X_{prompt}, LLaMA2-7B, acting as a "radiologist intern", generates a preliminary report draft x'_0. This draft is produced through instruction tuning by optimizing the cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{draft} &= \mathcal{L}(\theta_{draft}; X'_{draft}, X_{prompt}, V_I) \\ &= -\sum_{i=1}^{L} \log p_{\theta_{draft}}(x'_i | V_I, X_{prompt}, X_{draft,<i}),\end{aligned} \tag{5}$$

where θ_{draft} are the trainable parameters, $X_{draft,<i}$ denotes the preceding predicted tokens, and L is the sequence length.

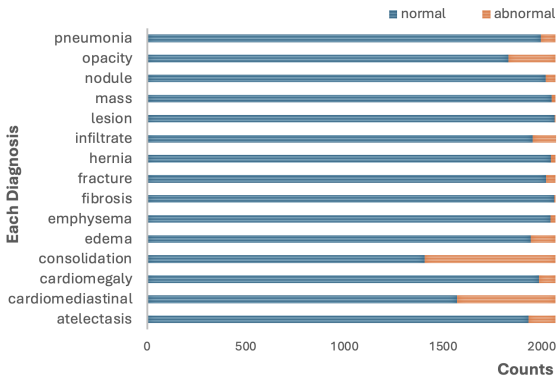


Figure 4: The normality distribution of each diagnosis across the whole dataset.

The generated draft serves dual purposes. During inference, it is used as the initial noise for the reverse diffusion process, replacing pure Gaussian noise to better align initial noise the training and inference stages. During the training stage, the draft is randomly substituted for the ground-truth report as the input vector before noise injection, ensuring diversity and robustness in the training process for the denoiser. This dual use of the draft enhances both the consistency and the performance of the diffusion-based framework.

3.3 Prior-Guided Conditioning

The effectiveness of conditioning techniques in enhancing content quality in diffusion models has been widely demonstrated across various domains, including image generation [14, 15, 34] and captioning tasks [2, 31]. In the context of radiology report generation, distinguishing between normal and abnormal cases is especially critical, as the focus and requirements of these two scenarios differ significantly. However, RRG datasets often exhibit a highly imbalanced normality distribution, as shown in Figure 4. This imbalance underscores the importance of incorporating techniques that explicitly account for and differentiate between normal and abnormal cases during the report generation process.

To achieve this, we introduce a binary domain-specific prior prompt that indicates the normality of each sample, derived from a predefined knowledge graph [10, 17]. This prior prompt acts as an additional conditioning signal for the diffusion process. Specifically, we utilize a prompt embedding layer, implemented as a multi-layer perceptron (MLP), to encode the prior prompt into the same dimensional space as the textual input embedding x_0. This allows for seamless addition-based information fusion. The encoded prompt

vector is then incorporated into the model by adding it to the textual input embedding immediately after the report embedding layer, effectively integrating domain-specific knowledge into the training process. The final textual input is defined as:

$$x_0 = x_0 + g_{pri}(x_{prior}), \tag{6}$$

where $g_{pri}(\cdot)$ represents the learnable embedding layer for the prior prompt, and x_{prior} denotes the binary prior information related to the sample's normality.

Though concatenation-based fusion could also be a viable alternative for DADNET, we opt for element-wise addition-based information fusion due to its alignment with the strength of diffusion models, particularly their efficiency. Addition operations are computationally faster and require fewer learnable parameters compared to concatenation layers [18, 46]. Furthermore, while concatenation-based approaches are often beneficial when preserving spatial information is essential, this consideration is not relevant for our text generation tasks, making addition a more efficient and suitable choice.

This prior-guided conditioning improves the model's ability to tailor its output based on the clinical characteristics of the input case, ensuring that the generated reports are accurate and contextually appropriate, while also maintaining computational efficiency.

The overall training objective for the proposed DADNET framework integrates multiple components to optimize both generation quality and diversity. The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{draft} + \mathcal{L}_{xe} + \mathcal{L}_{eos} + \mathcal{L}_{bit}. \tag{7}$$

4 Experiments

4.1 Experimental Settings

Datasets. Our experiments include the most widely-recognized IU-Xray dataset, a public RRG benchmark collected at Indiana University. This dataset comprises 3,955 radiology reports, each corresponding to a single case with associated frontal and/or lateral chest X-ray images, totaling 7,470 radiographs. The data is divided into training, validation, and testing sets in a 7:2:1 ratio. As part of preprocessing, all images are resized to dimensions of $3 \times 224 \times 224$ and normalized, while the reports are cleaned by removing punctuation and replacing infrequent words (appearing fewer than three times) with the placeholder token <unk>. In addition, the sequence length for each report is capped at a maximum of 60 words. The preprocessing aligns with the techniques outlined in COMG [10].

Evaluation Metrics. We evaluate DADNET from two complementary perspectivesto ensure a comprehensive assessment. First, we adopt standard natural language generation (NLG) metrics (i.e., BLEU[1-4] [35], METEOR [1], ROUGE-L [25], and CIDEr [42]) to measure the quality, fluency, and relevance of the generated reports compared to reference texts. Second, we assess the model's ability to produce diverse output using CIDEr [42] (capturing intra-sample diversity), Self-BLEU [49] (evaluating inter-sample diversity by assessing the dissimilarity of generated reports), and Div-4 [5] (measuring n-gram diversity). These metrics together provide a holistic view of DADNET's performance in balancing report quality and diversity.

Table 2: Diversity Analysis on IU-Xray

	Models	CIDEr ↑	Div-4 ↑	Self-BLEU ↓
AR	R2Gen†	0.566	0.839	0.363
AR	R2GenCMN†	0.609	0.873	0.306
NAR	D^2-Net†	0.389	0.939	0.195
NAR	DADNET	0.712	0.958	0.135

Implementation Details. We employ the pretrained ResNet101 [12] as the visual feature extractor and LLAMA2-7B [41] as the large language model. The diffusion transformer comprises an encoder and a decoder, each configured with 3 layers. For each multi-head attention block, the number of attention heads is set to 8. The reverse process during inference operates over 100 time steps. Model optimization is performed using the ADAM optimizer [20] with an initial learning rate of 5×10^-4, which is linearly warmed up over the first 20,000 training steps. The training procedure is conducted with a batch size of 16.

4.2 Results and Discussion

Quantitative Results. For NLG metrics, DADNET demonstrates superior performance over the baseline model, D^2-Net, across all evaluation metrics, benefiting from the integration of both contributions. Notably, there is a substantial improvement in CIDEr, which not only reflects the quality of the generated content but also captures its diversity. Given its dual significance, CIDEr is also included in our diversity evaluation metrics.

For NLG metrics, DADNET achieves superior performance over the baseline model, D^2-Net, across all evaluation metrics, driven by the integration of both contributions. Among these metrics, CIDEr demonstrates a notable improvement. As a metric designed to evaluate the quality of generated content by comparing it to reference texts while assigning greater weight to rare and informative n-grams, CIDEr is a particularly informative metrics for the RRG task. Its ability to highlight less frequent but diagnostically critical phrases makes it an effective measure of diversity in RRG. For instance, a model that generates diverse and detailed descriptions encompassing multiple regions of interest (ROIs) within a report achieves a higher CIDEr score compared to one producing repetitive or generic phrases. This is due to CIDEr's emphasis on the presence of diverse and meaningful n-grams, rewarding models that deliver rich and comprehensive outputs aligned with reference data. This dual role of CIDEr, capturing both content quality and diversity, underpins its inclusion in our diversity assessment.

In the diversity evaluation, Table 2 presents a comprehensive comparison between DADNET and several advanced RRG models, including two AR methods — R2Gen [4], R2GenCMN [3] — and one NAR baseline, D^2-Net [19]. As shown in the table, the diffusion-based framework exhibits a notable improvement in generating inter-sample diversity, as reflected in the metrics Self-BLEU and Div-4, which measure diversity across outputs for a single input sentence. Both DADNET and D^2-Net outperform AR methods in terms of Div-4 and Self-BLEU. However, D^2-Net appears to compromise intra-sample diversity, as evidenced by its lower CIDEr score. In contrast, DADNET achieves substantial improvements across all

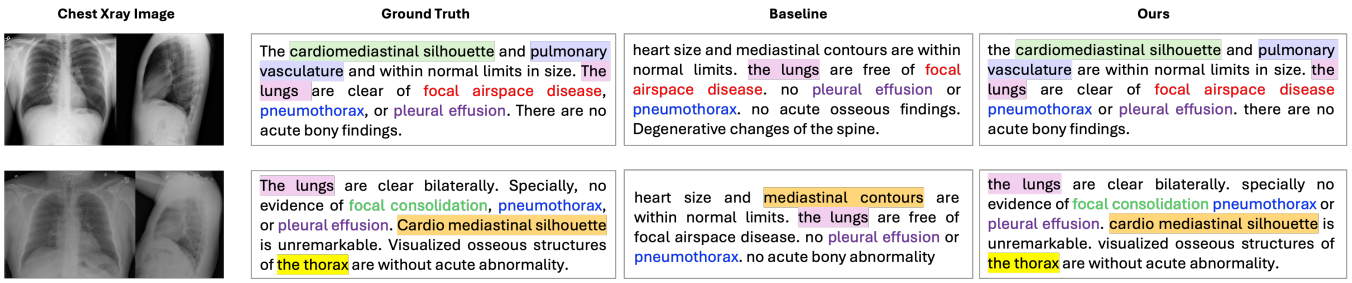


Figure 5: Qualitative analysis of the generated reports. Key organs are marked in corresponding colours, and diagnostic information is distinguished using varying text colours.

evaluation metrics, underscoring its capacity to balance intra- and inter-sample diversity effectively in radiology report generation. Notably, DADNET achieves a 0.103% increase in CIDEr, a 0.085% improvement in Div-4, and a 0.171% reduction in Self-BLEU relative to R2GenCMN, which is widely regarded as the baseline method in recent works.

Qualitative Results. To evaluate the effectiveness of DADNET from an alternative perspective, we conduct case studies on two selected samples, as shown in Figure 5. The reports generated by DADNET are compared against those produced by the baseline model, D^2-Net as well as the ground-truth. For clarity, key medical details are highlighted using distinct visual cues: different marking colors are used to represent specific organs or regions of interest (ROIs), while varied text colors denote specific diagnostic terms. From the figure, it is evident that the reports generated by DADNET exhibit superior quality compared to those from D^2-Net. Notably, DADNET successfully identifies all relevant ROIs, and diagnostic details within each report, whereas D^2-Net falls short in these aspects. This demonstrates that DADNET not only effectively detects ROIs in radiographs but also delivers precise diagnostic information.

Table 3: Ablation on Different Components

Base	LLM	Priors	B@3	B@4	M	Cr
✓			0.197	0.129	0.177	0.335
✓		✓	0.204	0.145	0.202	0.376
✓	✓		0.220	0.165	0.201	0.622
✓	✓	✓	0.225	0.175	0.203	0.708

Ablation Studies. Table 3 reports the results of our ablation studies on the IU-Xray dataset, showcasing the individual contributions and overall effectiveness of the proposed components: LLM-Assisted Noise Initialization and Prior-Guided Conditioning. The findings indicate that employing the prior-guided mechanism alone primarily enhances NLG metrics, reflecting an improvement in the quality of the generated reports. In contrast, integrating the LLM-assisted module significantly amplifies the model's performance, particularly in terms of diversity. These results emphasize the complementary roles of these components and their combined impact in achieving superior outcomes of DADNET.

5 Conclusion

In this paper, we propose DADNET, an innovative diffusion-based framework aimed at generating accurate and diverse radiology reports. By integrating large language models (LLMs), DADNET addresses the noise gap between the training and inference phases in diffusion-based generation models, enhancing the diversity of the generated content. Additionally, it incorporates prior knowledge prompts to embed domain-specific information, further supporting the report generation process. Our method achieves state-of-the-art performance on a widely recognized benchmark, demonstrating exceptional quality and both intra- and inter-sample diversity. In future work, we aim to explore the diversity dimension of RRG tasks within alternative frameworks to identify the optimal approach for producing medical reports that achieve a balance between high quality and diversity without compromising either.

References

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. https://aclanthology.org/W05-0909/

[2] Ting Chen, Ruixiang Zhang, and Geoffrey E. Hinton. 2023. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=3itjR9QxFw

[3] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5904–5914. https://doi.org/10.18653/V1/2021.ACL-LONG.459

[4] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1439–1449. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.112

[5] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. 2019. Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 10695–10704. https://doi.org/10.1109/CVPR.2019.01095

[6] Nicolas Devanthery, Natalie Heracleous, Benoit Dufour, Jean-Daniel Fardel, Benoit Rizk, Hugues Brat, Cyril Thouly, Henning Müller, Federica Zanca, and Lluis Borràs Ferrís. 2024. Structured radiology report text analysis using Natural Language Processing for automatic billing. In *Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*, Vol. 12929. SPIE, 204–208.

[7] Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. 2024. Exploiting the Signal-Leak Bias in Diffusion Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*. IEEE, 4013–4022. https://doi.org/10.1109/WACV57701.2024.00398

[8] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. 2023. Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 22873–22884. https://doi.org/10.1109/ICCV51070.2023.02096

[9] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=jQj-_rLVXsj

[10] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2024. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7995–8004.

[11] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, and Weidong Cai. 2025. ORID: Organ-Regional Information Driven Framework for Radiology Report Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[14] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.* 23 (2022), 47:1–47:33. https://jmlr.org/papers/v23/21-0635.html

[15] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022). https://doi.org/10.48550/ARXIV.2207.12598 arXiv:2207.12598

[16] Weijian Huang, Cheng Li, Hong-Yu Zhou, Hao Yang, Jiarun Liu, Yong Liang, Hairong Zheng, Shaoting Zhang, and Shanshan Wang. 2024. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications* 15, 1 (2024), 7620.

[17] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 19809–19818. https://doi.org/10.1109/CVPR52729.2023.01897

[18] Hatem Ibrahem, Ahmed Salem, and Hyun Soo Kang. 2022. Exploration of Semantic Label Decomposition and Dataset Size in Semantic Indoor Scenes Synthesis via Optimized Residual Generative Adversarial Networks. *Sensors* 22, 21 (2022), 8306. https://doi.org/10.3390/S22218306

[19] Yuda Jin, Weidong Chen, Yuanhe Tian, Yan Song, Chenggang Yan, and Zhendong Mao. 2024. Improving Radiology Report Generation with D 2-Net: When Diffusion Meets Discriminator. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2215–2219.

[20] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[21] Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. 2023. Improving Zero-shot Visual Question Answering via Large Language Models with Reasoning Question Prompts. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 4389–4400. https://doi.org/10.1145/3581783.3612389

[22] Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024. KARGEN: Knowledge-Enhanced Automated Radiology Report Generation Using Large Language Models. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 15005)*, Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (Eds.). Springer, 382–392. https://doi.org/10.1007/978-3-031-72086-4_36

[23] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. 2024. Enhancing Advanced Visual Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 1915–1929. https://aclanthology.org/2024.emnlp-main.114

[24] Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Tom Weidong Cai. 2024. Multimodal Causal Reasoning Benchmark: Challenging Vision Large Language Models to Infer Causal Links Between Siamese Images. *CoRR* abs/2408.08105 (2024). https://doi.org/10.48550/ARXIV.2408.08105 arXiv:2408.08105

[25] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[26] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024. Common Diffusion Noise Schedules and Sample Steps are Flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*. IEEE, 5392–5399. https://doi.org/10.1109/WACV57701.2024.00532

[27] Cheng Liu, Chao Wang, Yan Peng, and Zhixu Li. 2024. ZVQAF: Zero-shot visual question answering with feedback from large language models. *Neurocomputing* 580 (2024), 127505. https://doi.org/10.1016/J.NEUCOM.2024.127505

[28] Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based Multimodal Curriculum Learning for Medical Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3001–3012. https://doi.org/10.18653/V1/2021.ACL-LONG.234

[29] Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025. Interactive Evaluation for Medical LLMs via Task-oriented Dialogue System. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, 4871–4896. https://aclanthology.org/2025.coling-main.325/

[30] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 11451–11461. https://doi.org/10.1109/CVPR52688.2022.01117

[31] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023. Semantic-Conditional Diffusion Networks for Image Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 23359–23368. https://doi.org/10.1109/CVPR52729.2023.02237

[32] Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Fang Zeng, Xi Jiang, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, Dajiang Zhu, Dinggang Shen, Tianming Liu, and Xiang Li. 2024. An Iterative Optimizing Framework for Radiology Report Summarization With ChatGPT. *IEEE Trans. Artif. Intell.* 5, 8 (2024), 4163–4175. https://doi.org/10.1109/TAI.2024.3364586

[33] Xin Mei, Libin Yang, Denghong Gao, Xiaoyan Cai, Junwei Han, and Tianming Liu. 2024. PhraseAug: An Augmented Medical Report Generation Model With Phrasebook. *IEEE Trans. Medical Imaging* 43, 12 (2024), 4211–4223. https://doi.org/10.1109/TMI.2024.3416190

[34] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8162–8171. http://proceedings.mlr.press/v139/nichol21a.html

[35] 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, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[36] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance. *CoRR* abs/2311.18681 (2023). https://doi.org/10.48550/ARXIV.2311.18681 arXiv:2311.18681

[37] Han Qin and Yan Song. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 448–458. https://doi.org/10.18653/V1/2022.FINDINGS-ACL.38

[38] Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2023. Protocol: Large language model-based information extraction from free-text radiology reports: a scoping review protocol. *BMJ open* 13, 12 (2023).

[39] Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digit. Medicine* 7, 1 (2024). https://doi.org/10.1038/S41746-024-01219-0

[40] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 7433–7442. https://doi.org/10.1109/CVPR52729.2023.00718

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). https://doi.org/10.48550/ARXIV.2302.13971 arXiv:2302.13971

[42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

[43] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2GenGPT: Radiology Report Generation with Frozen LLMs. *CoRR* abs/2309.09812 (2023). https://doi.org/10.48550/ARXIV.2309.09812 arXiv:2309.09812

[44] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. 2021. A Self-Boosting Framework for Automated Radiographic Report Generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2433–2442. https://doi.org/10.1109/CVPR46437.2021.00246

[45] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. 2024. FreeInit: Bridging Initialization Gap in Video Diffusion Models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 15061)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 378–394. https://doi.org/10.1007/978-3-031-72646-0_22

[46] Haodong Yan, Zijun Liu, Jinglong Chen, Yong Feng, and Jun Wang. 2023. Memory-augmented skip-connected autoencoder for unsupervised anomaly detection of rocket engines with multi-source fusion. *ISA Transactions* 133 (2023), 53–65. https://doi.org/10.1016/j.isatra.2022.07.014

[47] Yueting Yang, Xintong Zhang, Jinan Xu, and Wenjuan Han. 2024. Empowering Vision-Language Models for Reasoning Ability through Large Language Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 10056–10060. https://doi.org/10.1109/ICASSP48485.2024.10446407

[48] Yi Zhou, Lei Huang, Tao Zhou, Huazhu Fu, and Ling Shao. 2021. Visual-Textual Attentive Semantic Consistency for Medical Report Generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 3965–3974. https://doi.org/10.1109/ICCV48922.2021.00395

[49] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1097–1100. https://doi.org/10.1145/3209978.3210080