

Diffusion Networks with Task-Specific Noise Control for Radiology Report Generation

Anonymous Authors

ABSTRACT

Existing radiology report generation (RRG) studies mostly adopt autoregressive (AR) approaches to produce textual descriptions token-by-token for specific clinical radiographs, where they are susceptible to error propagation problems if irrelevant contents are half-way generated, leading to potential ill-presenting of precise diagnoses, especially when there exist complicated abnormalities in radiographs. Although the non-AR paradigm, e.g., diffusion model, provides an alternative solution to tackle the problem from AR by generating all contents in parallel, the mechanism of using Gaussian noise in existing diffusion models still has significant room to improve when such models are used in particular circumstances, i.e., providing proper guidance in controlling noises in the diffusive process to ensure precise report generation. In this paper, we propose to conduct RRG with diffusion networks by controlling the noise with task-specific features, which leverages irrelevant visual and textual information as noise rather than the stochastic Gaussian noise, and allows the diffusion networks to filter particular information through iterative denoising, thus performing a precise and controlled report generation process. Experiments on IU X-RAY and MIMIC-CXR demonstrate the superiority of our approach compared to strong baselines and state-of-the-art solutions. Human evaluation and noise type analysis show that comprehensive noise control greatly helps diffusion networks to refine the generation of global and local report contents.¹

CCS CONCEPTS

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

KEYWORDS

Radiology Report Generation, Diffusion Networks, Noise Control, Task-specific Noise

1 INTRODUCTION

Medical imaging holds a crucial position in clinical medicine and treatment guidance, where physicians are always required to write medical reports based on the syndromes depicted in images and thus create comprehensive professional records for patient references and later processes. As a particular category of medical

¹Code will be released in the final version of the paper.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or professional use, not for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnn>

images, radiographs play a vital role in assessing patients' health by examining the internal structures of their bodies, and have been widely used in cardiology, dentistry, and pulmonology, etc. Generally, writing reports is a time-consuming job and often error-prone for inexperienced radiologists, which thus drives a series of work [3, 16, 20, 39, 45] on generating reports automatically and precisely. These studies achieve significant success on this topic, proving the feasibility of this research direction.

To effectively generate radiology reports, most existing studies [4, 13, 20, 28, 35, 36, 50, 52, 53] leverage autoregressive (AR) models (e.g., LSTM [11] and Transformer [46]) as their foundation architecture with the encoder-decoder pipeline. In doing so, visual encoders are jointly optimized with text decoders to capture essential semantics from radiograph inputs so as to establish well image-text mapping for the generation process. They normally adopt latent representations to store the semantic information for such mapping and there is a potential deficiency where those representations have ambiguities in conveying all essential abnormalities in the radiograph. The text decoder is thus disrupted by representation noises and has difficulties in generating comprehensive reports. Moreover, the AR-based text decoder has its own problem in susceptibility to error propagation, thus potentially generates contextually incoherent diagnoses if irrelevant contents are half-way produced.

With the recent advances of non-AR paradigm, e.g., diffusion model [9] on text generation [7, 24] and other cross-modal scenarios [2], it thus provides an alternative solution to existing AR-based approaches for RRG. However, in applying diffusion models, it is difficult to perform precise RRG with directly using stochastic Gaussian noises, thus it still requires certain task-specific guidance to help handle the necessary information and smooth the generation process. Although some studies [16, 28, 34, 53] have shown effectiveness by leveraging external medical knowledge as guidance for AR-based approaches, their designs are not easy to be applied to diffusion networks, whose integrity of information optimization is likely to be corrupted. Therefore, an effective approach is expected to enhance diffusion networks for RRG.

In this paper, we propose a non-AR solution for RRG with diffusion networks, namely, CONTROLDIFF, by employing a novel task-specific noise control mechanism to appropriately operate essential cross-modal information in noising and denoising processes. In our approach, we distinguish useful features, e.g., visual representations of different radiographs and textual contents in reports, from others and leverage the non-useful information as the noise in our diffusion networks, rather than the standard stochastic Gaussian one. Particularly, for each modality, we process both global and local information from them to construct our noise vectors, where removing global noises improves the coherence of final reports and removing local ones enhances the accuracy of describing specific regions in input radiographs. Experimental results on two benchmark datasets, i.e., IU X-RAY and MIMIC-CXR, demonstrate the

117 superiority of our approach against state-of-the-art studies. Hu- 175
 118 man evaluation and quantitative noise type analysis illustrate that 176
 119 choosing different noises affects content filtering during iterative 177
 120 generation, where controlling global noises ensures the overall 178
 121 consistency of the generated reports, and controlling local noise 179
 122 provides fine-grained task-specific guidance for diffusion networks 180
 123 to produce precise reports. 181
 124

125 2 RELATED WORK 182

126 2.1 Radiology Report Generation 183

127 RRG is a domain-specific extension to image description generation 184
 128 [32]. This task requires to automatically generate reports in the 185
 129 medical domain. To perform the task, existing studies generally 186
 130 follow the encoding-decoding paradigm. They leverage a visual 187
 131 encoder to capture visual features and use a text decoder to produce 188
 132 reports, where advanced architectures such as convolutional neural 189
 133 networks and Transformers are used. To improve the performance 190
 134 of RRG, there are studies that try to identify essential visual and 191
 135 textual features that contribute to the task and leverage them ac- 192
 136 cordingly. These studies leverage regional visual features [22, 45], 193
 137 medical terms [18, 53, 54], knowledge graphs [13, 28, 56], and re- 194
 138 port templates [20, 25] to generate high-quality reports. There are 195
 139 other studies that put emphasis on improving the cross-modal align- 196
 140 ment through attention mechanisms [16, 29, 57], memory networks 197
 141 [3, 39, 48], expert tokens [50], etc. These studies achieve promising 198
 142 performance for RRG, while they mainly rely on the AR paradigm 199
 143 and thus suffer from error propagation issues. Compared with these 200
 144 studies, the approach proposed in this paper is based on diffusion 201
 145 networks, which generate tokens at the same time and thus avoid 202
 146 error propagation issues. 203
 147
 148

149 2.2 Diffusion Networks 204

150 Diffusion models [9] are non-auto-regressive approaches that are 205
 151 widely used for image generation. Recently, the diffusion models 206
 152 and their variants have been applied to text generation tasks and 207
 153 demonstrated as an alternative solution with outstanding perfor- 208
 154 mance on cross-modal content generation [19, 42, 43]. Owing to 209
 155 the discrete nature of texts, several studies that use diffusion net- 210
 156 works propose to model discrete data with continuous forms, e.g., 211
 157 embedding [7, 24] and bit representations [2, 31]. For example, Li 212
 158 et al. [24] propose diffusion-based networks for text generation by 213
 159 projecting discrete tokens into continuous vectors. Chen et al. [2] 214
 160 model discrete texts with binary bit representations and enhance 215
 161 the text generation process with a self-condition mechanism for 216
 162 image captioning. These studies generally propose new model ar- 217
 163 chitecture or utilize new features to improve the denoising process 218
 164 of diffusion models, so as to generate high-quality reports, where 219
 165 the standard stochastic Gaussian noise is used to in the noising 220
 166 process. Compared to these aforementioned studies, our approach 221
 167 utilizes task-specific noise rather than stochastic Gaussian noise to 222
 168 control diffusion networks for RRG. 223
 169
 170

171 3 THE APPROACH 224

172 Given an input radiograph \mathcal{V} , our approach generates its corre- 225
 173 sponding radiology report $\hat{\mathcal{R}}$ following the pipeline shown in Figure 226
 174

1 with three main components, namely, the visual encoder, the task- 175
 specific noise generator (TNG), and the diffusion networks (DN). 176
 Specifically, the visual encoder f_{VE} encodes the input radiograph 177
 \mathcal{V} into visual representations \mathbf{v} . The TNG f_{TNG} provides a noise 178
 vector \mathbf{n} for the diffusion networks f_{DN} through two main com- 179
 ponents, i.e., the global noise generator (GNG) and the local noise 180
 generator (LNG). GNG uses background Visual features shared by 181
 most radiographs and non-informative n-grams in the reports to 182
 construct the global noise vector \mathbf{n}_G ; LNC leverages regional visual 183
 features of undetected regions in \mathcal{V} and irrelevant medical terms 184
 of reports to construct the local noise vector \mathbf{n}_L . Then, GNC and 185
 LNC fuse \mathbf{n}^G and \mathbf{n}^L into the noise vector \mathbf{n} that is processed by 186
 f_{DN} afterwards. Finally, DN utilizes \mathbf{v} and \mathbf{n} to produce $\hat{\mathcal{R}}$. In the 187
 following texts, we illustrate the details of each aforementioned 188
 component according to the pipeline sequence. 189
 190

191 3.1 Visual Encoder 192

193 The visual encoder aims to encode the input radiograph \mathcal{V} into 194
 a latent representation \mathbf{v} . It contains two components, namely, a 195
 visual feature extractor f_{VE} and a feature encoder f_{FE} . Specifically, 196
 f_{VE} is a pre-trained vision backbone model (i.e., ResNet-101 [8]), and 197
 f_{FE} follows the standard architecture of Transformer [46] encoder. 198
 We first adopt f_{VE} to extract visual features \mathbf{h}^v from \mathcal{V} and obtain 199
 \mathbf{h}^v from the last convolutional layer of f_{VE} through 200

$$201 \mathbf{h}^v = f_{VE}(\mathcal{I}) \quad (1) \quad 202$$

203 Then, we employ f_{FE} to encode \mathbf{h}^v into the visual representations 204
 \mathbf{v} of \mathcal{V} through 205

$$206 \mathbf{v} = f_{FE}(\mathbf{h}^v) \quad (2) \quad 207$$

208 where \mathbf{v} is used in TNG to produce noise vectors and DN for report 209
 generation. 210

211 3.2 Task-specific Noise Generator 212

213 Generally, diffusion networks utilize the stochastic Gaussian noise 214
 [2, 10, 23, 55] and present promising results in text generation tasks. 215
 Since the noise is not relevant to any task information, it is intuitive 216
 to explore whether task-related noise is able to improve model 217
 performance. Particularly for RRG, the forward noising process is 218
 analogized to adding non-essential information that is not relevant 219
 to the abnormalities highlighted in the gold standard report \mathcal{R}^* ; the 220
 denoising process is to eliminate such information from a noise text 221
 to reproduce \mathcal{R}^* . We propose to control the noise of diffusion net- 222
 works with task-specific characteristics. Specifically, we consider 223
 two types of noise, namely global noise and local noise. We pro- 224
 pose global noise generator (GNG) and local noise generator (LNG) 225
 in leveraging global and local information from radiographs and 226
 reports to provide the two types of task-specific noise for diffusion 227
 networks. Details of GNG and LNG are illustrated as follows. 228

229 *Global Noise Generator.* The GNC constructs the global noise 230
 vector \mathbf{n}_G according to the visual and text global task-specific noise 231
 information that is shared by most radiographs and reports. Specif- 232
 ically, for global visual noise information, it is intuitive to regard 233
 the background features shared by the most radiographs as the 234
 noise. We run an off-the-shelf background segmentation toolkit 235
 (e.g., OpenCV [5]) to produce the background feature vector based 236

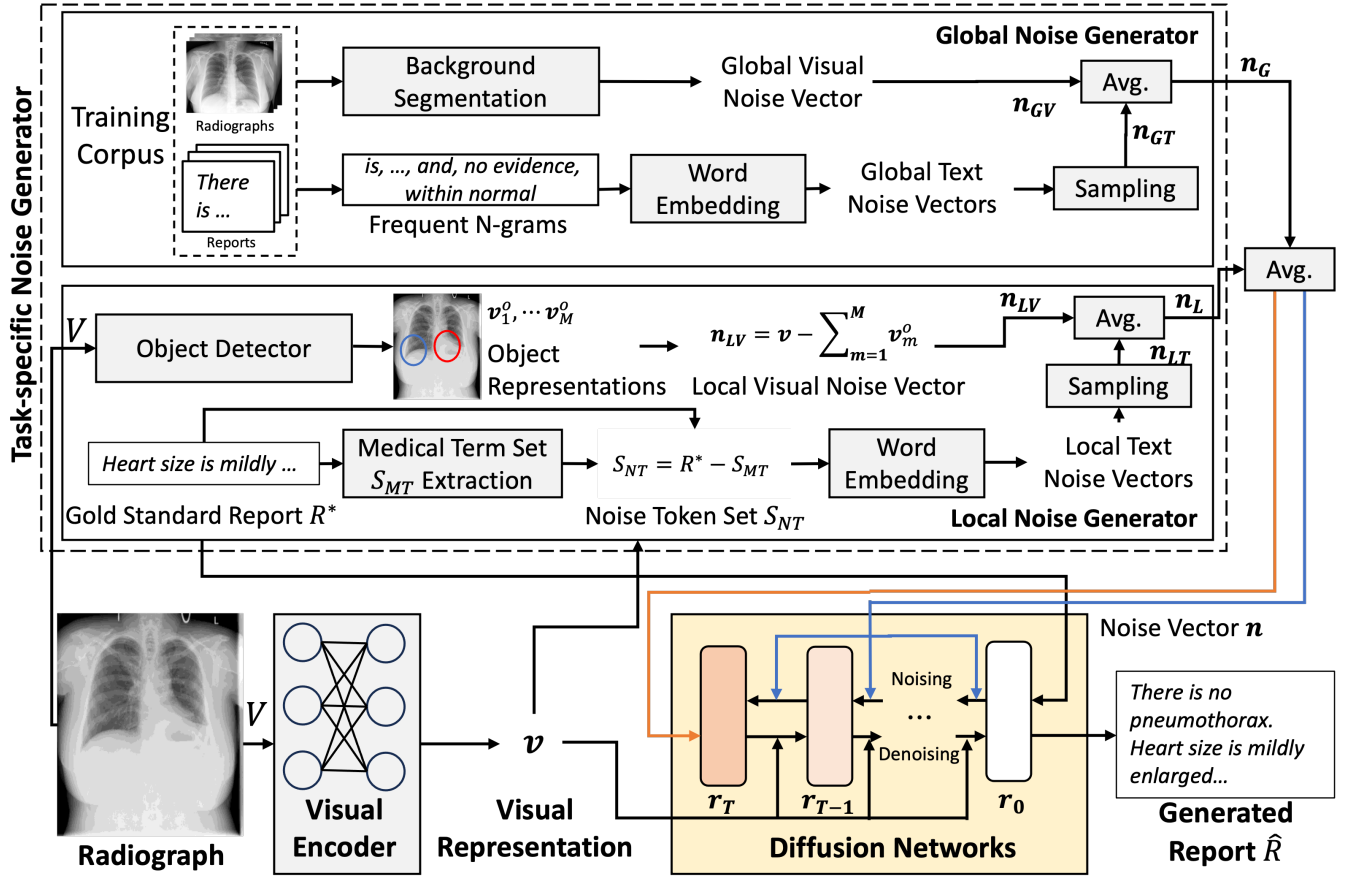


Figure 1: The overview architecture of our approach for RRG. It consists of three main components, namely, the visual encoder, the task-specific noise generator, and the diffusion networks, which are presented at the left-bottom, top, and right-bottom of the figure, respectively. The blue and orange arrows illustrate how the noise vector n is used in training and inference, respectively. We present an example radiograph for better demonstration.

on the radiographs in the entire training set and regard it as the global visual noise vector n_{GV} .

For global text noise information, we utilize n-grams that frequently appear in most reports, since they could be interpreted as stop words or report templates that carry limited information about the abnormalities in the radiograph. Specifically, we compute the frequencies of used n-grams in all radiology reports from the training set and select the top- N_{GT} ones as report templates. Then, we map the n-grams into their embeddings and regard them as the global text noise vectors $n_{GT,1} \dots n_{GT,N_{GT}}$. We randomly sample a vector from $n_{GT,1} \dots n_{GT,N_{GT}}$ and regard it as the global ext noise vector n_{GT} .

Finally, the global noise vector n^g is obtained with the average and normalization (*Norm*) of n^{gv} and n^{gt} through

$$n_G = \text{Norm} \left(\frac{1}{2} (n_{GV} + n_{GT}) \right) \quad (3)$$

Local Noise Generator. The LNC constructs the local noise vector n_L with fine-grained information in the radiographs \mathcal{V} and the gold standard radiology reports \mathcal{R}^* . We consider two types of noise from the visual and textual perspectives. For the visual noise, we use

irrelevant regional information in \mathcal{V} to construct the local visual noise vector n_{LV} . Specifically, we employ an off-the-shelf object detector (i.e., fine-tuned Fast R-CNN [40] on Chest Imagenome [51]) to extract regional visual features v^r from \mathcal{V} . We decompose v^r into $\{v_1^r \dots v_M^r\}$ along the channel dimension with M as the number of resulting representation. Considering the detected objects are generally essential regions in the radiograph, we use the overall representation obtained from Eq. (2) to subtract the regional features to compute the local visual noise vector n_{LV} .

$$n_{LV} = v - \sum_{m=1}^M (v_m^r) \quad (4)$$

For the text noise, we notice that radiology reports generally contain medical terms (e.g., *heart*, *lungs*, etc.) that are essential for analyzing the diseases of the patients. This motivates us to use non-medical-term words in \mathcal{R}^* to construct the local textual noise vector n_{LT} . Specifically, we firstly train a Transformer-based model to annotate medical terms and use it to extract the medical term set S_{MT} from the gold standard radiology report \mathcal{R}^* . Then, we regard the tokens in \mathcal{R}^* yet not in S_{MT} as the noise tokens. Similar to the

process to obtain global text noise, we use the same approach to map the noise tokens into embeddings, randomly sample one from the noise token embeddings, and use it as the local text noise vector \mathbf{n}_{GT} . Similar to the process in GNG, we compute \mathbf{n}_L according to \mathbf{n}_{LV} and \mathbf{n}_{LT} through

$$\mathbf{n}_L = \text{Norm}(\mathbf{n}_{LV} + \mathbf{n}_{LT}) \quad (5)$$

Once \mathbf{n}_G and \mathbf{n}_L are obtained, we compute the noise vector \mathbf{n} for the diffusion networks in the following processes by

$$\mathbf{n} = \text{Norm}(\mathbf{n}_G + \mathbf{n}_L) \quad (6)$$

3.3 Diffusion Networks

The DN (f_{DN}) aims to generate the final report $\widehat{\mathcal{R}}$ based on \mathbf{v} and \mathbf{n} . It consists of the diffusion noising and denoising processes, where both processes are used in training, and only the denoising process is used in inference. The following text illustrates the details of training and inference.

Training. In training, diffusion noising firstly adds the noise vector \mathbf{n} from TNG into the representation \mathbf{r}_0 of the gold standard report \mathcal{R}^* and obtain the noisy representations \mathbf{r}_t at the t -th step. The time step t is randomly sampled from a uniformed distribution $U(0, T)$ with T denoting the total number of steps. We follow the approach in DDCap [58] to convert tokens of \mathcal{R}^* into the one-hot representation and compute the representation \mathbf{r}_t at the t -th step with \mathbf{n} through

$$\mathbf{r}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{r}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \mathbf{n} \quad (7)$$

where $\bar{\alpha}_t$ is a blending scalar correlated to the noise scheduling strategy of denoising diffusion probabilistic model (DDPM) [9]. Then, f_{DN} reconstructs \mathbf{r}_t to \mathbf{r}_0 based on \mathbf{v} , where we compute the loss \mathcal{L} through

$$\mathcal{L} = \mathbb{E}_{t \sim U(0, T)} \|f_{DN}(\mathbf{r}_t, \mathbf{v}, t) - \mathbf{r}_0\|_2^2 \quad (8)$$

The trainable parameters in the model are updated accordingly through gradient descent.

Inference. Diffusion denoising generates $\widehat{\mathcal{R}}$ following the standard process of DDCap. It is worth noting that the process to obtain local text noise vector \mathbf{n}_{LN} relies on the gold standard radiology report \mathcal{R}^* , which is not available during inference. To handle this issue, we firstly collect a set \mathcal{S}_{All} with all tokens and a set \mathcal{S}_{MT} with all medical terms in the reports of the training data. Next, we randomly sample noise tokens from the difference of \mathcal{S}_{All} and \mathcal{S}_{MT} (i.e., $\mathcal{S}_{All} - \mathcal{S}_{MT}$) and following the same process in training to get the local text noise vector. Then, we denoise \mathbf{n} into the final representation $\widehat{\mathbf{r}}_0$. We initialize $\widehat{\mathbf{r}}_T$ with \mathbf{n} and iteratively subtract noises from $\widehat{\mathbf{y}}_T$ through

$$\widehat{\mathbf{r}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \frac{\widehat{\mathbf{r}}_t - \sqrt{1 - \bar{\alpha}_t} \cdot f_{DN}(\widehat{\mathbf{r}}_t, \mathbf{v}, t)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1 - \bar{\alpha}_{t-1}} \cdot f_{DN}(\widehat{\mathbf{r}}_t, \mathbf{v}, t) \quad (9)$$

Finally, we convert the one-hot representation $\widehat{\mathbf{r}}_0$ into tokens according to the vocabulary and obtain the final radiology report, which is denoted as $\widehat{\mathcal{R}}$.

Table 1: The statistics of IU X-RAY and MIMIC-CXR, where the numbers of images, reports, patients, and the token-based averaged length (AVG. LEN.) of reports in training, validation, and test sets are presented.

DATASET	IU X-RAY			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

4 EXPERIMENT

4.1 Experiment Settings

Datasets. We conduct our experiments on two conventional benchmark datasets, i.e., IU X-RAY [6] from Indiana University and MIMIC-CXR [17] from the Beth Israel Deaconess Medical Center. Table 1 reports the statistics of all datasets in terms of the numbers of radiographs, reports, patients, and word-based report length according to each split of the datasets. Specifically, IU X-RAY is relatively small with 7,470 chest X-Ray images and 3,955 radiology reports. MIMIC-CXR is the largest public radiology dataset with 473,057 chest X-ray images and 206,563 reports. We follow the convention of previous studies [3, 4, 13, 16, 39] by only preserving the ‘‘Findings’’ sections for both datasets. We use the dataset split with the ratio of 7:1:2 in Jing et al. [16] for IU X-RAY and the official split of MIMIC-CXR.

Baselines. To evaluate our proposed approach, we compare it with two baseline models in our experiments, including ‘‘TRANS’’ and ‘‘DIFF’’. ‘‘TRANS’’ represents the autoregressive baseline model with ResNet-101 [8] and a three-layer Transformer encoder as the visual encoder, and a three-layer Transformer decoder with an additional eight-head cross-attention layer as the decoder. ‘‘DIFF’’ is our baseline model with diffusion networks, which follows the same architecture as DDCap [58] and leverages stochastic Gaussian noise to control the diffusion networks. ‘‘DIFF+TNG’’ represents the model where ‘‘DIFF’’ is equipped with the proposed task-specific noise control in our approach, denoting our full model.

Evaluation. For evaluation metrics, we follow existing RRG studies [3, 4, 13, 21, 39] to evaluate the generated reports with two types of metrics, namely, natural language generation (NLG) and clinical efficacy (CE) metrics. NLG metrics measure the quality of generated reports based on n-gram overlapping, consisting of BLEU [38], METEOR [33] and ROUGE-L [26]. CE metrics evaluate the accuracy of estimating specific medical observations based on the following procedures. First, we adopt CheXpert [14]² to extract medical labels from both generated and gold standard reports. Then, we calculate the precision, recall, and F1 scores between the

²CheXpert annotates 14 categories of terms related to thoracic diseases and support devices, including *atelectasis*, *cardiomegaly*, *consolidation*, *edema*, *enlarged cardiom*, *fracture*, *lung lesion*, *lung opacity*, *no finding*, *pneumonia*, *pneumothorax*, *pleural effusion*, *pleural other*, and *support devices*.

Table 2: The guideline for human evaluation of the reports.

METRIC	SCORES	ILLUSTRATION
FLUENCY	1	The report is ungrammatical and hard to understand.
	2	The report has some grammatical issues but it is understandable.
	3	The report is grammatical and understandable.
COMPLETENESS	1	The report misses more than two essential abnormalities in the radiograph.
	2	The report misses one or two essential abnormalities in the radiograph.
	3	The report covers all essential abnormalities in the radiograph.
PRECISION	1	The report contains more than two incorrect descriptions of the abnormalities.
	2	The report contains one or two incorrect descriptions of the abnormalities.
	3	The descriptions of the abnormalities in the report are all correct.

Table 3: Performance (i.e., the average and standard deviation of three runs with different random seeds) of baselines and our approach (i.e., “+TNG”) on the test sets of IU X-RAY and MIMIC-CXR datasets in terms of NLG and CE metrics. We report both the average and standard deviation of three runs with different random seeds. BL-1, BL-2, BL-3, and BL-4 denote BLEU scores using uni-gram, bi-gram, tri-gram, and 4-grams; MTR and RG-L denote METEOR and ROUGE-L, respectively. The average improvement over all NLG metrics compared to “TRANS” is also presented in the “Avg. Δ” column. The relative improvements of our approach over baselines are statistically significant at $p \leq 0.05$ level.

DATA	MODEL	NLG METRICS							CE METRICS		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	Avg. Δ	P	R	F1
IU X-RAY	TRANS	0.385±0.003	0.219±0.002	0.150±0.002	0.105±0.003	0.150±0.002	0.305±0.003	-	-	-	-
	DIFF	0.414±0.004	0.245±0.003	0.162±0.004	0.109±0.002	0.162±0.003	0.312±0.002	6.9%	-	-	-
	+TNG	0.508±0.003	0.332±0.002	0.243±0.003	0.189±0.002	0.207±0.002	0.390±0.003	48.6%	-	-	-
MIMIC-CXR	TRANS	0.355±0.003	0.213±0.004	0.138±0.002	0.088±0.003	0.126±0.003	0.269±0.002	-	0.348±0.002	0.314±0.003	0.330±0.002
	DIFF	0.373±0.002	0.217±0.003	0.142±0.002	0.101±0.004	0.134±0.002	0.274±0.003	5.5%	0.385±0.003	0.401±0.004	0.393±0.003
	+TNG	0.411±0.004	0.265±0.002	0.183±0.002	0.132±0.003	0.186±0.002	0.299±0.004	30.3%	0.477±0.004	0.484±0.004	0.480±0.004

aforementioned obtained labels, and use the computed scores as results of the CE metrics.

In addition to automatic evaluations, we also perform human evaluations of the quality of the generated reports. For each report, we ask three annotators with medical backgrounds to assess its quality in three aspects: fluency, completeness, and precision. The guideline is illustrated in Table 2. The annotators are asked to rate each aspect of the report on a scale of 1 to 3 accordingly, with higher scores indicating better quality. The quality of a report is measured by the average scores from different annotators.

Implementation Details. For model architecture, we use the standard Transformer encoder with three multi-head attention layers as f_{FE} , and adopt two different eight-layer Transformers for f_{VE} in TNG and f_{DE} , respectively. The number of the attention head and dimension of the hidden states for all modules are set to 8 and 512, respectively. For the diffusion networks, the total time step for diffusion forwarding and decoding processes T is set to 20. We also follow the standard process of the denoising diffusion implicit model (DDIM) [44] sampler in the decoding process. For optimization, we use Adam optimizer to update all model parameters with a learning rate of $5e - 4$. We follow the learning rate scheduling strategy in Vaswani et al. [46] with 20,000 steps for warm-up, and train the model on IU X-RAY and MIMIC-CXR with 300 and 10 epochs,

Table 4: Human evaluation results on the report generated by different models using the sampled test instances from MIMIC-CXR. The range of human evaluation scores is from 1 to 3. “F”, “C”, and “P” denote fluency, completeness, and precision, respectively. “IAA” means the inter-annotator agreement (i.e., the number of scores agreed by all annotators out of all annotations).

DATA	MODEL	F	C	P	Avg.	IAA
IU X-RAY	TRANS	2.6	2.1	2.0	2.2	82%
	DIFF	2.7	2.2	2.0	2.3	80%
	DIFF+TNG	2.8	2.4	2.1	2.4	82%
MIMIC-CXR	TRANS	2.5	1.9	1.8	2.1	84%
	DIFF	2.5	2.0	1.9	2.1	82%
	DIFF+TNG	2.7	2.2	2.0	2.3	86%

respectively. For other hyper-parameter settings, we try different combinations of them and select the one with the best performance on the validation set in our final experiments. For all experiments, we run them three times with different random seeds and report the average and standard deviation.

Table 5: Performance comparison of our approach with the state-of-the-art studies on test sets of IU X-RAY and MIMIC-CXR with respect to NLG and CE metrics. The best results of different metrics are highlighted in boldface. For LLM-based approaches (i.e., XRAYGPT), we illustrate the number of parameters in parentheses.

DATA	MODEL	NLG METRICS						CE METRICS		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-RAY	ST [47]	0.216	0.124	0.087	0.066	-	0.306	-	-	-
	ATT2IN [41]	0.224	0.129	0.089	0.068	-	0.308	-	-	-
	ADAATT [30]	0.220	0.127	0.089	0.068	-	0.308	-	-	-
	CoATT [16]	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	HRGR [25]	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	CMAS-RL [15]	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	R2GEN [4]	0.470	0.304	0.219	0.165	-	0.371	-	-	-
	CA [29]	0.492	0.314	0.222	0.169	0.193	0.381	-	-	-
	CMCL [27]	0.477	0.305	0.217	0.162	0.186	0.378	-	-	-
	PPKED [28]	0.483	0.315	0.224	0.168	-	0.376	-	-	-
	R2GENCMN [3]	0.475	0.309	0.222	0.170	0.191	0.375	-	-	-
	R2GENRL [39]	0.494	0.321	0.235	0.181	0.201	0.384	-	-	-
	XRAYGPT (7B) [37]	0.177	0.104	0.047	0.007	0.105	0.203	-	-	-
	CONTROLDIFF	0.508	0.332	0.243	0.189	0.207	0.390	-	-	-
MIMIC -CXR	ST [47]	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
	ATT2IN [41]	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	ADAATT [30]	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
	TOPDOWN [1]	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	R2GEN [4]	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	CA [29]	0.350	0.219	0.152	0.109	0.151	0.283	-	-	-
	CMCL [27]	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
	PPKED [28]	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	R2GENCMN [3]	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	R2GENRL [39]	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
	WARMSTART [34]	0.392	0.245	0.169	0.124	0.153	0.285	0.359	0.412	0.384
	ITA [49]	0.395	0.253	0.170	0.121	0.147	0.284	-	-	-
	WARMSTART [34]	0.392	0.245	0.169	0.124	0.153	0.285	0.359	0.412	0.384
RGRG [45]	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447	
ORGAN [12]	0.407	0.256	0.172	0.123	0.162	0.293	0.416	0.418	0.385	
KiUT [13]	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321	
XRAYGPT (7B) [37]	0.128	0.045	0.014	0.004	0.079	0.111	-	-	-	
	CONTROLDIFF	0.411	0.265	0.183	0.132	0.186	0.299	0.477	0.484	0.480

4.2 Overall Results

Experiment results of different models on the two benchmark datasets are reported in Table 3, with several observations drawn as follows. It is observed that the basic non-AR model (“DIFF”) consistently outperforms the AR one (“TRANS”) on both datasets, where the reason owes to that the error propagation problem is alleviated by diffusion networks through synchronous generation. On top of “DIFF”, our full model “DIFF+TNG” obtains further improvements through leveraging task-specific noise rather than stochastic Gaussian noise in diffusion networks, confirming the effectiveness of noise control. The possible reason behind this observation is that the task-specific noise provides more precise hints to the diffusion process, therefore ensuring the quality of generated reports by eliminating potential irrelevant contents.

For human evaluation, we randomly sample 50 instances from the test sets of IU X-RAY and MIMIC-CXR and collect the reports generated by different models (i.e., “TRANS”, “DIFF”, “DIFF+TNG”). The results of the human evaluation are reported in Table 4. We also report the inter-annotator agreement (IAA) that computes the number of scores agreed by all annotators out of all annotations. Similar to the trend in Table 3, human evaluation results show that our approach outperforms all baselines, which further confirms the effectiveness of our approach.

Moreover, we compare it with existing state-of-the-art solutions on both datasets, with results presented in Table 5. Overall, our approach outperforms other AR-based solutions on all metrics, illustrating the superiority of our approach for RRG. Notably, our approach even achieves better performance than the studies that based on large language models (LLMs) (i.e., XRAYGPT), indicating

Table 6: Performance (i.e., the average and standard deviation of three runs with different random seeds) comparison of our approach under different settings of noise control on IU X-RAY and MIMIC-CXR with respect to NLG metrics. “GN” means the diffusion model with the standard stochastic Gaussian noise; “GVN” and “GTN” are global visual noise and global text noise, respectively; “LVN” and “LTN” are local visual noise and local text noise, respectively. “✓” means the noise type is used in the model; the last row where “GVN”, “GTN”, “LVN”, and “LTN” are used is our full model.

(a) IU X-RAY											
ID	NOISE TYPE					EVALUATION METRIC					
	GN	GVN	GTN	LVN	LTN	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
1	✓					0.414±0.004	0.245±0.003	0.162±0.004	0.109±0.002	0.162±0.003	0.312±0.002
2		✓				0.445±0.003	0.267±0.003	0.188±0.002	0.130±0.004	0.168±0.003	0.342±0.002
3			✓			0.442±0.002	0.265±0.002	0.185±0.004	0.132±0.004	0.166±0.002	0.340±0.003
4				✓		0.443±0.002	0.264±0.003	0.189±0.004	0.134±0.002	0.167±0.002	0.341±0.003
5					✓	0.447±0.002	0.268±0.002	0.190±0.003	0.134±0.004	0.170±0.003	0.344±0.003
6		✓	✓			0.458±0.002	0.308±0.003	0.216±0.002	0.156±0.004	0.189±0.003	0.367±0.002
7				✓	✓	0.460±0.003	0.305±0.004	0.218±0.002	0.159±0.003	0.188±0.003	0.369±0.003
8		✓		✓		0.463±0.003	0.313±0.002	0.217±0.003	0.158±0.002	0.186±0.003	0.371±0.004
9			✓		✓	0.462±0.003	0.311±0.002	0.216±0.002	0.157±0.002	0.187±0.003	0.370±0.004
10	✓	✓	✓	✓	✓	0.505±0.002	0.330±0.003	0.242±0.004	0.186±0.002	0.206±0.002	0.391±0.002
11		✓	✓	✓	✓	0.508±0.003	0.332±0.002	0.243±0.003	0.189±0.002	0.207±0.002	0.390±0.003

(b) MIMIC-CXR											
ID	NOISE TYPE					EVALUATION METRIC					
	GN	GVN	GTN	LVN	LTN	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
1	✓					0.373±0.002	0.217±0.003	0.142±0.002	0.101±0.004	0.134±0.002	0.274±0.003
2		✓				0.387±0.003	0.229±0.002	0.156±0.004	0.114±0.003	0.151±0.002	0.279±0.003
3			✓			0.385±0.003	0.230±0.004	0.153±0.002	0.109±0.002	0.148±0.004	0.280±0.003
4				✓		0.383±0.002	0.234±0.003	0.157±0.003	0.110±0.004	0.146±0.003	0.281±0.002
5					✓	0.384±0.003	0.232±0.004	0.155±0.002	0.111±0.003	0.150±0.004	0.283±0.002
6		✓	✓			0.397±0.004	0.251±0.004	0.168±0.002	0.126±0.003	0.169±0.002	0.285±0.003
7				✓	✓	0.391±0.003	0.250±0.004	0.170±0.002	0.124±0.003	0.168±0.002	0.289±0.003
8		✓		✓		0.401±0.003	0.254±0.004	0.172±0.003	0.129±0.004	0.172±0.003	0.287±0.003
9			✓		✓	0.395±0.004	0.253±0.003	0.174±0.003	0.126±0.004	0.171±0.002	0.292±0.003
10	✓	✓	✓	✓	✓	0.408±0.002	0.262±0.003	0.180±0.002	0.132±0.004	0.187±0.002	0.302±0.003
11		✓	✓	✓	✓	0.411±0.004	0.265±0.002	0.183±0.002	0.132±0.003	0.186±0.002	0.299±0.004

that appropriate modeling of the report generation process is more efficient than using massive parameters in LLMs.

4.3 Effect of Different Noise Types

To explore the impact of controlling different noise types, we run experiments on using particular task-specific information as noise in diffusion networks. Table 6 reports the results on two benchmark datasets, where “GN” refers to Gaussian noise; “GVN” and “GTN” represent the global visual and textual noise, respectively; “LVN” and “LTN” denote the local visual and textual noise, respectively. Several observations from different perspectives are illustrated as follows. First, compared with the model with the standard Gaussian noise (i.e., ID=1), our approach (ID=2-11) with any type of task-specific noise achieves better performance on most evaluation metrics. This observation demonstrates the effectiveness of using task-related noise in diffusion models for improving model performance. Second, comparing our approach with global noise

(ID=6) and with local noise (ID=7), the models achieve similar performance, showing both global and local noise contribute to the task; similarly, our approaches with visual (ID=8) and text noise (ID=9) obtain similar performance, showing both visual and text features are essential to the task. Third, we observe that models with multi-modal features as noise (ID=6-9) obtain improvements over the ones using single-modal features (ID=2-5), since controlling visual- or textual-only noise provides coarse guidance for report generation. Finally, we observe that our approaches with (ID=10) and without (ID=11) the standard Gaussian noise achieve comparable performance. This observation is intuitive since the Gaussian noise is task-irrelevant and thus brings limited useful information compared with other task-specific types of noise. Furthermore, our approach obtains the best performance, where leveraging various noise information from different views and modalities refines the iterative generation process.

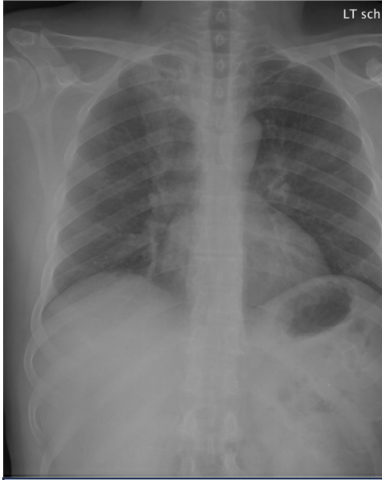
	TRANS
	<p>The image shows a well-defined bone structure, and there seems to be a blur on the top left, possibly indicating a fracture. There's a darker area at the center which might be an artifact. It is hard to determine the state of the internal organs. A shadow on the left side could suggest a foreign object presence.</p>
	DIFF
	<p>The chest X-ray reveals clear bone outlines, and there is an increased density visible in the upper section which might imply a broken rib. The heart shape appears regular. An unclear patch towards the lower left might be a processing error.</p>
	DIFF+GNG
	DIFF+LNG
Gold Standard	DIFF+LNG
<p>There is XXXX increased opacity within the right upper lobe with possible mass and associated area of atelectasis or focal consolidation. The cardiac silhouette is within normal limits. XXXX opacity in the left midlung overlying the posterior left 5th rib may represent focal airspace disease. No pleural effusion or pneumothorax. No acute bone abnormality.</p>	<p>There is increased opacity observed in the right upper lung field which could suggest a mass, along with an adjacent region that might represent atelectasis or focal consolidation. The cardiac outline seems normal. The left lung demonstrates an area of density that may correspond with airspace disease, although less clearly defined.</p>
	DIFF+TNG (Ours)
	<p>There is increased opacity within the right upper lobe, potentially indicating a mass and an associated area of atelectasis or focal consolidation. The cardiac silhouette is within normal limits. Obscured opacity in the left midlung may suggest focal airspace disease. There is no evidence of pleural effusion or pneumothorax. Bone structures appear normal without acute abnormalities.</p>

Figure 2: An illustration of the report generation processes (through texts generated at different time steps) by different models with an example radiograph. Medical terms shared by the model outputs and the gold standard texts are highlighted in the same color. “GNG” and “LNG” stand for global noise generator and local noise generator, respectively.

4.4 Case Study

To further qualitatively demonstrate the effect of noise control in our approach, we present a case study in Figure 2 with an example input radiograph selected from the test set of MIMIC-CXR, as well as the reports generated by “TRANS”, “DIFF”, “DIFF+GNG”, “DIFF+LNG” and “DIFF+TNG”, where “DIFF+GNG” and “DIFF+LNG” refer to the diffusion networks equipped with only GNG or LNG, respectively, and medical terms shared by model output and gold standard are highlighted in same colors. It is observed that “DIFF” generates reports with more medical terms related to the input radiograph than “TRANS” since the error propagation problem of AR models is alleviated by “DIFF”. Meanwhile, the reports generated by “DIFF” still contain irrelevant descriptions, since a less controlled generation process is performed owing to the use of stochastic Gaussian noise. “DIFF+GNG” and “DIFF+LNG” improve the quality of generated reports compared to “DIFF”. “DIFF+GNG” effectively eliminates irrelevant descriptions by controlling global noises; “DIFF+LNG” offers more fine-grained noise control for the diffusion networks and produces reports with more related medical terms than “DIFF+GNG”. Finally, “DIFF+TNG” obtains the most elaborated reports compared to all aforementioned models, suggesting

that controlling both global and local noise provides comprehensive information for diffusion networks to generate precise reports.

5 CONCLUSION

In this paper, we propose CONTROLDIFF that utilizes diffusion networks to generate the report for RRG and thus does not suffer from the error propagation issues of the existing approaches that use AR models. In addition, we enhance the diffusion networks with task-specific noise (e.g., global and local visual and text features) rather than the standard stochastic Gaussian noise used in the standard diffusion networks, to generate precise reports for RRG. Experimental results on two widely used English benchmark RRG datasets, namely, IU X-RAY and MIMIC-CXR, indicate the superiority and effectiveness of our proposed approach compared to existing studies, where our approach outperforms strong baselines and existing studies on both datasets. Further analyses and case study explore the effect of our noise controlling mechanisms from different perspectives, suggesting that our approach presents its potential of being a reference framework to conduct a controlled generation process for other related tasks in future studies.

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*. 6077–6086.
- [2] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. 2023. Analog Bits: Generating Discrete Data using Diffusion Models with Self-conditioning. In *ICLR*. 1–23.
- [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 (Volume 1: Long Papers)*. Online, 5904–5914.
- [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)*. Online, 1439–1449.
- [5] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapov, and Mario Cifrek. 2012. A Brief Introduction to OpenCV. In *2012 proceedings of the 35th international convention MIPRO*. IEEE, 1725–1730.
- [6] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing A Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Medical Informatics Assoc.* 23, 2 (2016), 304–310.
- [7] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *ICLR*. 1–20.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas, NV, USA) (CVPR '16)*. 770–778.
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *NeurIPS* 33 (2020), 6840–6851.
- [10] Jonathan Ho and Tim Salimans. 2022. Classifier-free Diffusion Guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [12] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada, 8108–8122.
- [13] 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.
- [14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Iltus, Christopher Chute, Henrik Marklund, Behzad Haghighi, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 590–597.
- [15] Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, 6570–6580.
- [16] Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, 2577–2586.
- [17] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR: A De-identified Publicly Available Database of Chest Radiographs with Free-text Reports. *Scientific Data* 6 (2019).
- [18] Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Jadhav. 2023. Replace and Report: NLP Assisted Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada, 10731–10742.
- [19] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*. 1–17.
- [20] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 6666–6673.
- [21] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang. 2023. Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA, 3334–3343.
- [22] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2022. Auxiliary Signal-guided Knowledge Encoder-decoder for Medical Report Generation. *World Wide Web* (2022), 1–18.
- [23] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. *Advances in Neural Information Processing Systems* 35 (2022), 4328–4343.
- [24] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). 1–16.
- [25] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 1537–1547.
- [26] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain, 74–81.
- [27] 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 (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Online, 3001–3012.
- [28] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021*. 13753–13762.
- [29] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive Attention for Automatic Chest X-ray Report Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online, 269–280.
- [30] Jiaseen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. IEEE Computer Society, 3242–3250.
- [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.
- [32] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). 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.). 1–17.
- [33] Denkowski Michael and Lavie Alon. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, 85–91.
- [34] Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving Chest X-ray Report Generation by Leveraging Warm Starting. *Artificial Intelligence in Medicine* 144 (2023), 102633.
- [35] Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. 2022. Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates, 7123–7138.
- [36] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive Transformer-Based Generation of Radiology Reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Punta Cana, Dominican Republic, 2824–2832.
- [37] Sahal Shaji Mullapilly Hisham Cholakkal Rao Muhammad Anwer Salman Khan Jorma Laaksonen Omkar Thawkar, Abdelrahman Shaker and Fahad Shahbaz Khan. 2023. XrayGPT: Chest Radiographs Summarization using Large Medical Vision-language Models. *arXiv: 2306.07971* (2023).
- [38] 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*. Philadelphia, Pennsylvania, USA, 311–318.
- [39] Han Qin and Yan Song. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Dublin, Ireland, 448–458.

929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044

- 1045 [40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.
- 1046
- 1047 [41] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1179–1195.
- 1048
- 1049
- 1050 [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. 10684–10695.
- 1051
- 1052 [43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic Text-to-image Diffusion Models with Deep Language Understanding. *NeurIPS* 35 (2022), 36479–36494.
- 1053
- 1054 [44] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. 1–20.
- 1055
- 1056 [45] 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*. 7433–7442.
- 1057
- 1058 [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in neural information processing systems*. 5998–6008.
- 1059
- 1060 [47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 3156–3164.
- 1061
- 1062 [48] Jun Wang, Abhir Bhalerao, and Yulan He. 2022. *Cross-Modal Prototype Driven Network for Radiology Report Generation*. Lecture Notes in Computer Science, Vol. 13695. 563–579.
- 1063
- 1064 [49] Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. 2022. An Inclusive Task-Aware Framework for Radiology Report Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (Eds.)*. Cham, 568–577.
- 1065
- 1066 [50] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 11558–11567.
- 1067
- 1068
- 1069
- 1070
- 1071
- 1072
- 1073
- 1074
- 1075
- 1076
- 1077
- 1078
- 1079
- 1080
- 1081
- 1082
- 1083
- 1084
- 1085
- 1086
- 1087
- 1088
- 1089
- 1090
- 1091
- 1092
- 1093
- 1094
- 1095
- 1096
- 1097
- 1098
- 1099
- 1100
- 1101
- 1102
- [51] Joy T. Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest Imaging Genome Dataset for Clinical Reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). 1–14.
- [52] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Punta Cana, Dominican Republic, 4009–4015.
- [53] Jingyi You, Dongyuan Li, Manabu Okumura, and Kenji Suzuki. 2022. JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). 5989–6001.
- [54] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment. *ArXiv abs/1907.09085* (2019).
- [55] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. DiffuSum: Generation Enhanced Extractive Summarization with Diffusion. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada, 13089–13100.
- [56] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. When Radiology Report Generation Meets Knowledge Graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. 12910–12917.
- [57] 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)*. 3965–3974.
- [58] Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. 2022. Exploring Discrete Diffusion Models for Image Captioning. *arXiv preprint arXiv:2211.11694* (2022).
- 1103
- 1104
- 1105
- 1106
- 1107
- 1108
- 1109
- 1110
- 1111
- 1112
- 1113
- 1114
- 1115
- 1116
- 1117
- 1118
- 1119
- 1120
- 1121
- 1122
- 1123
- 1124
- 1125
- 1126
- 1127
- 1128
- 1129
- 1130
- 1131
- 1132
- 1133
- 1134
- 1135
- 1136
- 1137
- 1138
- 1139
- 1140
- 1141
- 1142
- 1143
- 1144
- 1145
- 1146
- 1147
- 1148
- 1149
- 1150
- 1151
- 1152
- 1153
- 1154
- 1155
- 1156
- 1157
- 1158
- 1159
- 1160