Article

Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation

Daniel Parres [1,*], Alberto Albiol [1] and Roberto Paredes [1,2]

[1] Campus de Vera, Universitat Politècnica València, Camí de Vera s/n, 46022 Valencia, Spain; alalbiol@iteam.upv.es (A.A.); rparedes@prhlt.upv.es (R.P.)
[2] Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera s/n, 46022 Valencia, Spain
* Correspondence: dparres@prhlt.upv.es

Abstract: Deep learning is revolutionizing radiology report generation (RRG) with the adoption of vision encoder–decoder (VED) frameworks, which transform radiographs into detailed medical reports. Traditional methods, however, often generate reports of limited diversity and struggle with generalization. Our research introduces reinforcement learning and text augmentation to tackle these issues, significantly improving report quality and variability. By employing RadGraph as a reward metric and innovating in text augmentation, we surpass existing benchmarks like BLEU4, ROUGE-L, F_1CheXbert, and RadGraph, setting new standards for report accuracy and diversity on MIMIC-CXR and Open-i datasets. Our VED model achieves F_1-scores of 66.2 for CheXbert and 37.8 for RadGraph on the MIMIC-CXR dataset, and 54.7 and 45.6, respectively, on Open-i. These outcomes represent a significant breakthrough in the RRG field. The findings and implementation of the proposed approach, aimed at enhancing diagnostic precision and radiological interpretations in clinical settings, are publicly available on GitHub to encourage further advancements in the field.

Keywords: radiology report generation; reinforcement learning; text augmentation; machine learning; deep learning; vision transformer; chest X-rays; medical image; text generation

check for updates

Citation: Parres, D.; Albiol, A.; Paredes, R. Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation. *Bioengineering* **2024**, *11*, 351. https://doi.org/10.3390/bioengineering11040351

Academic Editor: Simon W. Rabkin

Received: 12 March 2024
Revised: 29 March 2024
Accepted: 2 April 2024
Published: 3 April 2024

1. Introduction

Radiology report generation (RRG) is a challenging task where the goal is to interpret radiographic images and generate detailed reports on potential patient pathologies. In contrast to typical computer vision (CV) tasks, which aim to identify objects in images, RRG focuses on diagnosing potential pathologies and determining their presence, absence, or uncertainty. Moreover, limited data availability and the diverse nature of medical reports pose significant challenges. This task is crucial for streamlining and enhancing patient care, as it enables the analysis of individuals' health and the rapid detection of diseases. Furthermore, it serves as assistance and support for medical professionals. The RRG task parallels other works such as [1], where complex data are handled, and models trained from such data must ensure specific security criteria. This similarity arises from the need to develop models capable of diagnosing pathologies precisely, as in this type of problem, patients' health is put at risk. Current approaches mainly rely on deep learning (DL), specifically the vision encoder–decoder (VED) architecture [2–8], incorporating components like memories [9,10] or reinforcement learning (RL) [5,11,12] to improve performance.

This study presents a two-stage VED pure-transformer architecture for chest RRG. In the first stage, conventional negative log-likelihood (NLL) training is employed, while the second stage focuses on RL optimization for various metrics. These metrics encompass embedding comparison for semantic coherence [13], entity and relationship graph generation for pathology descriptions [14], and NLL. Additionally, we propose the integration of text augmentation and hard negative mining techniques. This comprehensive approach surpasses current chest RRG methodologies, enhancing report variability, factual correctness, completeness, and overall model generalization.

The key contributions of this study are as follows:

- Harnessing vision encoder–decoder transformer frameworks and reinforcement learning to enhance radiology report quality, factual correctness, and completeness.
- Proposing text augmentation to the training workflow. This technique allows for leveraging the scarcity of data by generating new reports. This leads to an improvement in diversity, averaging a threefold increase compared to the state of the art.
- Due to the challenge of measuring the quality of generated reports, we focus on employing specialized natural language processing-based metrics to evaluate radiology reports comprehensively.
- Open-sourcing our methodology on GitHub at https://github.com/dparres/Diversifying-Radiology-Report-Generation (accessed on 1 April 2024) to propel progress in diagnostic precision and radiological interpretation within clinical settings.

Related Work

In recent years, deep learning techniques have led various approaches to classify and detect pathologies in X-ray images. Notable efforts in this domain include Schlegl et al. [15], who proposed a convolutional network (CNN) for classifying tissue patterns in tomographies, utilizing semantic descriptions in reports as labels. Building on this success, subsequent neural network models were explored for X-rays, such as Shin et al. [16], who introduced a CNN for chest X-ray images and a recurrent network (RNN) for annotations, jointly trained to annotate diseases, anatomy, and severity. Other approaches, like that of Moradi et al. [17], focused on annotation through the concatenation of a CNN and an RNN block to identify regions of interest. Notably, Rubin et al. [18] employed parallelly trained CNNs for frontal and lateral chest X-ray views to estimate possible pathologies. The interest in RRG has grown, led by impactful models like TieNet [4], as well as innovative contributions from Li et al. [3], Jing et al. [19], and Jing et al. [19], each leveraging advanced techniques such as pretrained networks, coattention mechanisms, and retrieval policy modules for efficient disease classification and report generation.

Deep learning applications in radiology encounter obstacles due to limited and unstructured data availability, exacerbating the normal–abnormal case imbalance and complicated by ambiguous radiologist reports [20]. CNN models such as AlexNet [21], VGG-16 [22], GoogLeNet [23], and ResNet [24] dominate medical text–image mining, with the increasing prevalence of end-to-end training [20]. However, the model comparison is hindered by dataset diversity, although incorporating radiology reports is expected to grow [20]. Radiology reports, diverse in style and influenced by biases, pose challenges in training robust models, with few datasets meeting scalability and accessibility criteria [25]. Most RRG systems concentrate on X-ray tests, with emerging applications for CT scans and MRI datasets, each encountering unique challenges [25].

Alfarghaly et al. [26] introduced a novel technique for generating radiology reports from chest X-rays with DistilGPT2 [27]. This approach highlights faster training and improved metrics. Although limitations persist, larger datasets' release is suggested for enhanced model generalization and critical evaluation of quantitative methods [26]. Yang et al. [28] presented a method combining general medical and specific medical knowledge with a multihead attention mechanism for chest RRG. Furthermore, Pan et al. [29] proposed a method for generating chest radiology reports through cross-modal multiscale feature fusion. This architecture aims to enhance location sensitivity and disease characterization. Additionally, these advancements pave the way for aligning scales and integrating knowledge graphs, enhancing the accuracy of report generation [29]. Yang et al. [30] introduced a highly accurate and automated radiology generation framework coupled with a novel automatic medical knowledge updating mechanism, enhanced by a multimodal alignment approach. Zhao et al. [31] presented a knowledge enhancement technique leveraging medical knowledge in dictionary form alongside historical knowledge, complemented by a multilevel alignment method to mitigate modal differences between text and image. Nicolson et al. [32] introduced CvT2DistilGPT2, leveraging a convolutional vision transformer for optimal encoder warm starting and highlighting GPT2's [33] superiority in decoder warm starting over BERT [34]. Despite the proposals' novelty, they all rely on metrics such as BLEU and ROUGE-L for comparison, which may not be entirely suitable for measuring the presence, absence, or uncertainty of pathologies. Furthermore, these metrics fail to consider the interrelation of pathologies within the report. Additionally, the monotony and repetition in generated reports are noteworthy factors attributed to the inherent difficulty of the RRG task.

Due to the increasing interest in automating RRG, specific architectures have been designed to address chest RRG tasks. Liu et al. [5] proposed a CNN for extracting visual features, followed by a sentence decoder and word decoder for generating report topics and composing the final report. The model is fine-tuned using RL with CheXpert [35] labels. Chen et al. [9] introduced the memory-driven transformer, employing relational and memory-driven layers to enhance information retention and incorporation into the decoder. Meanwhile, Chen et al. [10] presented a VED based on cross-model memory networks, leveraging shared memory to capture and utilize visual–textual alignments.

Liu et al. [6] and Liu et al. [7] proposed leveraging unsupervised construction of knowledge graphs to replicate radiologists' patterns to generate reports. Focusing on RL, Miura et al. [11] suggested employing two metrics: one ensuring the generation of domain entities (fact$_{ENT}$) and the other maintaining coherent entity descriptions (fact$_{ENTNLI}$). These metrics are optimized alongside BERTScore [13], a semantic equivalence metric. Delbrouck et al. [12] propose a competitive chest RRG approach utilizing a VED. The model employs a DenseNet-121 [36] as the optical encoder and a single-layer BERT [34] as the decoder. Trained with RL, the optimization involves three metrics: RadGraph [14], BERTScore, and NLL. RadGraph, a neural network, constructs semantic annotations by forming a graph of entities and relationships in reports. The reward is computed by comparing the hypotheses' entities and relationships against the reference, resulting in higher-quality reports surpassing values achieved by fact$_{ENT}$ and fact$_{ENTNLI}$. These proposals employ chest-RRG-oriented metrics to enable more precise comparisons of report quality compared to metrics such as BLEU or ROUGE-L. Despite this and using RL for training, the reports exhibit repetitiveness and a lack of diversity.

2. Materials and Methods

This section of the paper presents the primary approach for the chest RRG task. Firstly, the neural models are introduced, and we propose four different architectures to analyze the most suitable for the RRG task. Subsequently, the training workflow of the models is outlined. Then, the text augmentation technique is introduced to leverage the scarcity of data. The final subsection presents the databases to be utilized and the metrics for comparing models and measuring the quality of the generated reports.

2.1. Proposed Architectures

Our model architectures for addressing the chest RRG problem are VED models, with Swin [37] as the vision model and BERT [34] as the language model or decoder. We opted for Swin for its outstanding performance in various computer vision tasks such as classification, object detection, and semantic segmentation [37]. Compared to ViT [38], Swin's hierarchical nature and more efficient attention mechanism based on shifted windows make it more suitable for X-ray analysis. Our models utilize both the base and small variants, denoted as SwinB and SwinS. These architectures are pretrained with ImageNet [39] and feature an input image size of three channels with a width and height of 224 pixels. They have architectural depths of 2, 2, 18, and 2 for each layer, with a patch size of 4. The primary difference lies in the number of heads, with SwinB utilizing sizes 4, 8, 16, and 32 and SwinS using 3, 6, 12, and 24.

For the decoder model, BERT was selected due to its proven effectiveness across various natural language processing (NLP) tasks [34,40,41]. However, we diverge from the standard configuration by employing only three layers with a hidden size of 1024, rather than the original 12 layers. Moreover, our investigation includes the analysis of two distinct decoders: one word-based, with a vocabulary of approximately 9.8k words, and another subword-based, featuring a vocabulary of 30k subwords. The integration of the vision model with the decoder model is facilitated through cross-attention layers, enabling the introduction of visual features extracted from the radiograph into the decoder. This integration allows for the autoregressive generation of reports.

Following this approach, the proposed VED models for chest RRG analysis are SwinB+BERT9k, SwinB+BERT30k, SwinS+BERT9k, and SwinS+BERT30k.

2.2. Training Workflows

In radiological contexts, multiple images (such as anterior–posterior and lateral views) are commonly generated per patient study. We employ a multi-image approach to address this, restricting it to three images per study. This involves concatenating features extracted by the Swin encoder for each image and processing them in the decoder. Furthermore, we apply random transformations—including translation, scaling, rotation, and adjustments to brightness and contrast—to augment training images.

Our training process comprises two stages, as illustrated in Figure 1. In the initial stage, the NLL loss function is employed, utilizing teacher forcing to train the model based on the generated hypothesis and the reference report. This stage typically encompasses approximately twelve training epochs. Subsequently, in the second stage, RL is introduced using the self-critical sequence training (SCST) algorithm [42]. SCST performs two forward passes of the VED: greedy decoding and beam-search multinomial sampling (sampling decoding). The first pass involves inference through greedy decoding to obtain the report Y_g without calculating gradients. Meanwhile, the second pass employs sampling decoding, calculating gradients to obtain the sampling report Y_s.

RL aims to optimize specific metrics by using them as rewards. In our approach, we propose utilizing two metrics: BERTScore [13] and RadGraph [14]. BERTScore effectively improves grammar and semantics within models, while $\mathrm{F}_1\mathrm{RG}_{ER}$ aids the model in prioritizing pathology-related entities and relationships.

After obtaining the reports Y_g and Y_s, we must calculate a loss for each metric to optimize ($Loss_{metric}$). This $Loss_{metric}$ is subsequently used to compute a weighted sum to obtain the final $Loss_{RL}$ and update the model weights. Equation (1) outlines the process to derive the $Loss_{metric}$, where the metric rewards are computed for Y_s and Y_g using the reference report Y_{ref}. These rewards are then subtracted and multiplied by the logarithm of the probabilities calculated during the generation of the Y_s report.

$$Loss_{metric}(Y_s, Y_g, Y_{ref}) = -(r_{metric}(Y_s, Y_{ref}) - r_{metric}(Y_g, Y_{ref}))\log(\Pr(Y_s)) \tag{1}$$

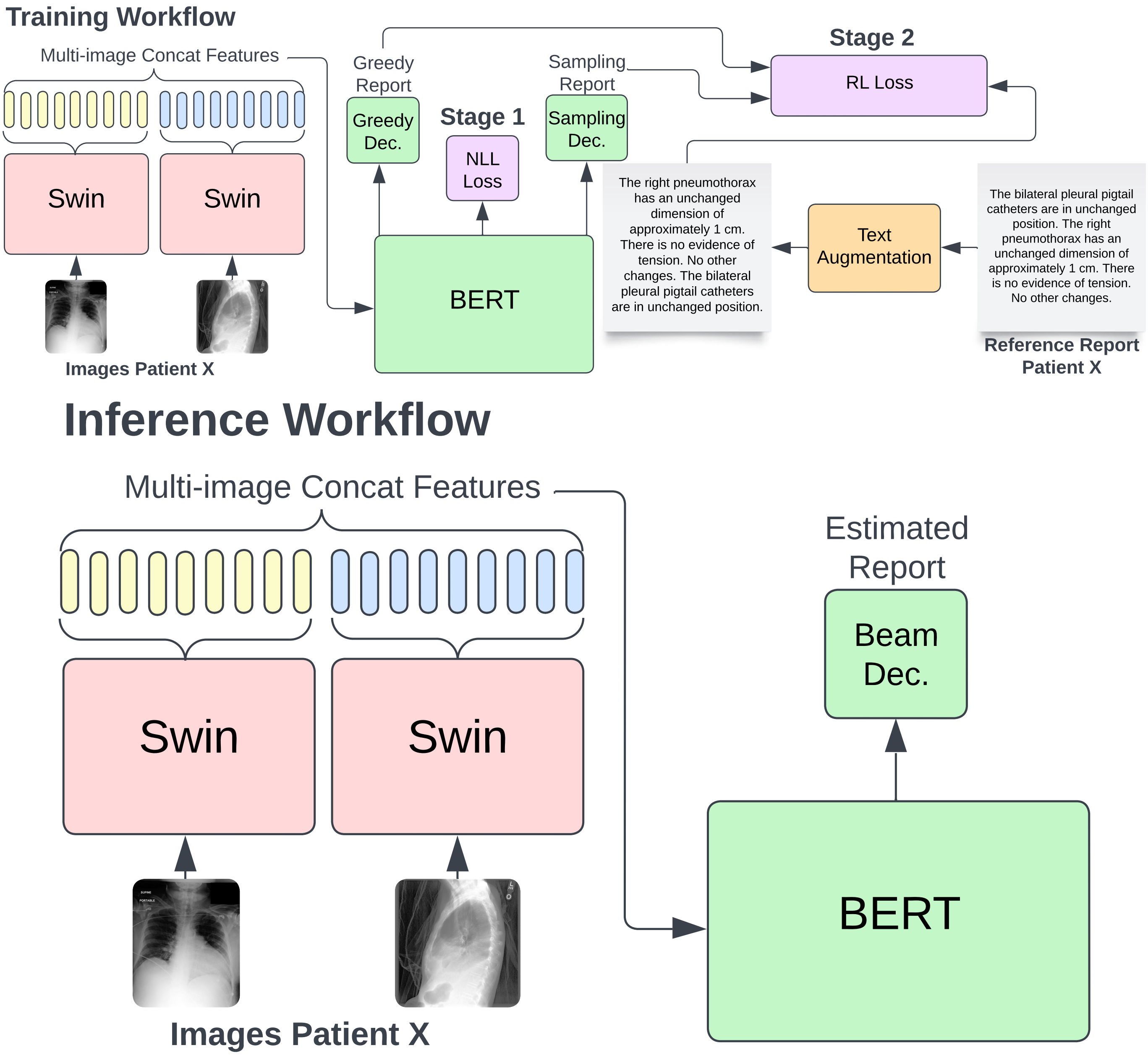
Each $Loss_{metric}$ contributes as a distinct weighted term to the final RL loss, as presented in Equation (2). To obtain the $Loss_{RL}$ for updating the model weights, $Loss_{BERTScore}$, $Loss_{\mathrm{F}_1 RG_{ER}}$, and the same loss utilized in the initial training stage, the $Loss_{NLL}$, are summed. In our experiments, we set $\alpha = \beta = 0.495$ and $\gamma = 0.010$ as the weighting factors.

$$Loss_{RL} = \alpha Loss_{BERTScore} + \beta Loss_{\mathrm{F}_1 RG_{ER}} + \gamma Loss_{NLL} \tag{2}$$

2.3. Text Augmentation

Data augmentation (DA) is the most prevalent technique for addressing machine learning problems with limited data. This approach significantly contributes to enhanced generalization and improved data utilization. However, regarding RRG, the focus has primarily been on augmenting input images. To the best of our knowledge, no prior approach has investigated the impact of augmenting radiology reports, a technique known as text augmentation (TA). In chest RRG, reports are often drafted by different individuals in diverse manners, lacking specific templates, which results in significant variability inherent to natural language. Presently, state-of-the-art RRG algorithms encounter this data scarcity issue, often generating highly repetitive and monotonous reports.

In this study, we introduce a TA technique designed to enhance the quality and diversity of generated reports. Our TA approach involves splitting the reference report into phrases, considering the typical structure of reports with multiple sentences. These phrases are then randomly reorganized to create new reports while maintaining the original diagnosis, as depicted in Figure 1. This method mitigates overfitting concerns and promotes a more targeted learning of the mentioned pathologies. Furthermore, our TA significantly improves the quality, factual accuracy, and comprehensiveness of the generated reports, as demonstrated subsequently.



The right pneumothorax has an unchanged dimension of approximately 1 cm. There is no evidence of tension. No other changes. The bilateral pleural pigtail catheters are in unchanged position.

The bilateral pleural pigtail catheters are in unchanged position. The right pneumothorax has an unchanged dimension of approximately 1 cm. There is no evidence of tension. No other changes.

Figure 1. Our training workflow using RL and TA to enhance report quality. In the initial stage, training with NLL employs teacher forcing. Subsequently, during RL training, the SCST algorithm computes rewards utilizing two different reports from two distinct decoding strategies: greedy search and beam-search multinomial sampling. Inference on validation and test sets is conducted using beam search in both stages.

The TA technique described and utilized in this work is exclusively employed in the training workflow, specifically during the NLL and RL stages. This technique aims to leverage the reference reports to enhance the model training process and increase its robustness. The report sentences will be randomly ordered in each training epoch for each training sample.

2.4. Datasets and Metrics

To evaluate the competitiveness of our models, we employed two publicly available datasets: MIMIC-CXR [43] and Open-i [44]. Due to the relatively small size of the Open-i dataset, containing 3.3 k reports, it was exclusively utilized for testing. The MIMIC-CXR dataset comprises 152 k reports for training, 2.3 k for validation, and 2.3 k for testing. The workflow involves training with the 152 k reports, saving the model weights based on the best performance on the validation set, and finally using these weights to perform inference on both the MIMIC-CXR and Open-i test sets. It is important to note that in this study, the models were specifically designed to generate the findings section of the reports. Additionally, samples with empty findings sections were excluded from consideration.

Metrics play a crucial role in comparing models, with BLEU4 [45] and ROUGE-L [46] currently considered as the most widely used natural language generation (NLG)-oriented metrics for reports. Additionally, two chest-RRG-oriented metrics, specifically designed to evaluate the quality of radiology reports, have been employed: F_1CheXbert (F_1cXb) [47] and F_1RG_{ER} [12,14]. These metrics leverage neural networks to assess the quality of generated reports, offering a higher semantic evaluation. F_1cXb utilizes CheXbert, a transformer model capable of identifying the presence of the 14 CheXpert [48] pathologies (atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pneumonia, pneumothorax, pleural other, support devices, and no finding) in a hypothesis report, and calculates the F_1-score based on the pathologies present in the reference report. On the other hand, F_1RG_{ER} employs a transformer model to analyze reports, generating graphs for entities and relationships. It evaluates comparisons between entity and relationship structures in hypotheses and reference graphs to compute the F_1-score.

3. Results

This section details the experiments carried out to develop models for chest RRG. We analyze the proposed architectures and then conduct an ablation study on effective training techniques and strategies. Finally, we assess the competitiveness of our proposal by comparing it with state-of-the-art models. All experiments were conducted using a computer equipped with an NVIDIA RTX 4090 GPU.

3.1. Analysis of Our Architectures

This study proposes four VED models, distinguished primarily by the number of trainable parameters in each component. Table 1 illustrates how SwinB+BERT30K is the largest model, while SwinS+BERT9k is the smallest. The remaining two models fall within a similar parameter range. Comparing architecture sizes is interesting because the RRG task operates in a data-limited environment due to the high variability of medical reports and their scarcity. Therefore, training models with high parameter counts can become challenging to tune efficiently. The number of parameters, along with the nature and size of the dataset, are crucial factors for effective learning that avoids underfitting and overfitting. Furthermore, the results presented in Table 1 encourage analysis of which components of the architectures may be more critical for obtaining meaningful reports.

In addition to the number of parameters, Table 1 showcases BLEU4, ROUGE-L, F_1cXb, and F_1RG_{ER} metrics. These metrics were obtained following the same workflow and hyperparameters across all four models. During the first stage, involving training with NLL, we conduct training for 12 epochs with a learning rate of 3×10^{-4}. Subsequently, in the second stage, corresponding to RL training, we train for 15 epochs with a learning rate of 5 × 10^{-5}. This learning rate value is crucial as it allows a good starting point for improving results based on previous training. The value of learning rates has been empirically explored using grid search with learning rate values of 5×10^{-3}, 3×10^{-3}, 1×10^{-3}, 5×10^{-4}, 3×10^{-4}, 1×10^{-4}, 5×10^{-5}, 3×10^{-5}, 1×10^{-5}, 5×10^{-6}, 3×10^{-6}, and 1×10^{-6} for the first and second training stages. Values above 5×10^{-5} in RL training lead to highly repetitive reports, while values below it yield only marginal improvements, failing to achieve significantly better results than training with NLL. The training epochs are set to 12 and 15 for the first and second stages, respectively, proving sufficient to achieve satisfactory results. Increasing the epochs does not significantly improve report quality due to the task's data scarcity and variability. During both stages, the learning rate is linearly decreased until reaching a value of zero at the end of the epochs. The results presented in Table 1 do not utilize image data augmentation, hard negative mining (HNM), or text augmentation.

Table 1. Architecture comparison using four report quality metrics on the MIMIC-CXR and Open-i test sets, RL training hours per epoch, and the number of parameters. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

Model	F_1cXb	F_1RG_{ER}	BLEU4	ROUGE-L	h/epoch	# Params
MIMIC-CXR						
SwinS+BERT9k RL	61.1	34.9	11.5	25.9	**12**	109 M
SwinS+BERT30k RL	61.4	34.3	11.3	25.2	24	130 M
SwinB+BERT9k RL	**63.1**	**35.4**	**11.7**	**26.5**	14	147 M
SwinB+BERT30k RL	62.9	35.1	11.6	**26.5**	30	168 M
Open-i						
SwinS+BERT9k RL	52.8	44.5	14.8	33.7	-	109 M
SwinS+BERT30k RL	52.8	44.1	14.6	33.4	-	130 M
SwinB+BERT9k RL	**54.7**	**45.6**	**15.1**	**34.3**	-	147 M
SwinB+BERT30k RL	54.4	45.3	14.9	34.1	-	168 M

Regarding report metrics on MIMIC-CXR, the SwinB+BERT9k model stands out, exhibiting superior performance compared to its closest counterpart, SwinB+BERT30k, and achieving a training time reduction of 16 h per epoch. For the SwinS models, results demonstrate a similarity, albeit with the BERT9k variant showcasing marginally better metrics across all parameters except F_1cXb. Interestingly, F_1cXb remains nearly identical between the two SwinS and SwinB models, indicating a direct impact of the encoder's size on this metric's competitiveness. Conversely, the choice between word-based or subword-based decoders does influence metrics such as BLEU4, ROUGE-L, and F_1RG_{ER}. Moreover, adopting word-based models yields superior metrics and significantly reduces training time, enhancing overall efficiency. The results from Open-i also rank the SwinB+BERT9k model as the most competitive and SwinS+BERT30k as the least competitive. Once again, this reaffirms that word-based models yield better results in the chest RRG task. Furthermore, the SwinS model demonstrates superior capability in generating reports by obtaining the best representations of radiographs.

Based on the results of the analysis, SwinB+BERT9k emerges as the most competitive and efficient model for chest RRG. However, various strategies can prove crucial in data-limited environments like chest RRG. This study specifically proposes using TA to leverage reports and significantly increase the training data. However, employing image augmentation techniques such as slight rotations, scale changes, shifts, random crops, and adjustments of brightness and contrast helps exploit the number of images. Moreover, since most datasets typically include more challenging samples than others, HNM is an option to consider. During training, samples with errors greater than the mean error plus their standard deviation are reutilized before moving to the next epoch.

Table 2 presents the ablation study of our best model, considering these strategies. The first row showcases the metrics achieved at the end of the NLL stage. The second row demonstrates how training with RL succeeds in improving these metrics; while image data augmentation shows a minor impact attributed to the robustness of transformers, HNM marginally enhances results by revisiting challenging samples at the end of each epoch. In contrast, TA emerges as a crucial technique, significantly improving chest-RRG-oriented metrics and BLEU4 by over one point. This underscores the significance and effectiveness of TA in efficiently leveraging the number of reports and furnishing the network with enhanced generalization capabilities.

Table 2. Ablation study of our best model on the MIMIC-CXR test set. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

Model	F_1cXb	F_1RG_{ER}	BLEU4	ROUGE-L
SwinB+BERT9k NLL	52.3	22.4	10.1	20.5
+ RL	63.1	35.4	11.7	26.5
+ Image Augment.	63.3	35.7	11.9	26.6
+ Hard Neg. Mining	64.2	35.8	12.0	26.8
+ Text Augment.	**66.2**	**37.8**	**13.2**	**27.1**

3.2. Benchmarking with the State of the Art

To assess the competitiveness of our top model against other state-of-the-art models, we selected the proposals with the most competitive results on both the MIMIC-CXR and Open-i datasets, as detailed in Table 3. The table is divided into two sections: models trained with NLL and models trained with RL, comparing them across the same metrics presented in Table 2. Metrics optimized for models trained with RL are indicated in parentheses. The benchmark model we employ is SwinB+BERT9k utilizing RL and TA, as depicted in Table 2.

The results underscore the importance of utilizing RL for training, as models trained with NLL exhibit inferior performance. In NLL models for MIMIC-CXR, it can be observed that Nicolson et al. [32]'s proposal achieves the highest BLEU4 score while securing a second position in terms of ROUGE-L. Pan et al. [29]'s proposal obtains the highest ROUGE-L value. Regarding chest RRG metrics, due to their novelty, not all proposals have registered values, as image-captioning metrics such as BLEU4 or ROUGE-L are commonly used. Our proposal achieved the highest F_1cXb values, followed by Delbrouck et al. [12]'s model; this trend was also observed for F_1RG_{ER} values. When compared to the NLL models of the Open-i dataset, the highest BLEU4 score is achieved by Miura et al. [11], while Chen et al. [10] obtain the highest ROUGE-L score. Similar to MIMIC-CXR, in chest-RRG-oriented metrics, our model obtains the highest values, followed by Delbrouck et al. [12].

Table 3 also presents the results for methods trained with RL. The metric used within the RL reward is indicated in parentheses. In the MIMIC-CXR dataset, it is apparent that the BLEU4 score reported by Miura et al. [11] deteriorates in both models when compared to the one trained with NLL. However, it still manages to enhance the F_1cXb score. Regarding Delbrouck et al. [12]'s models, there is an improvement of about one point in BLEU4 and ROUGE-L, with their model achieving the second-best scores for F_1cXb and F_1RG_{ER} at 62.2 and 34.7, respectively. Meanwhile, our SwinB+BERT9k model surpasses all RL models across all metrics, demonstrating superior quality, factual correctness, and completeness. In the case of the Open-i dataset, a similar trend is observed, with our model outperforming other proposals by approximately nine and six points in F_1cXb and F_1RG_{ER}, respectively, compared to Delbrouck et al. [12]. The good results obtained in this database demonstrate the generalization and adaptability capacity of our training workflow to the chest RRG task.

Upon analyzing the results in Table 3, it becomes evident that the conventional approach to compare RRG systems relies on image-captioning metrics such as BLEU4 and ROUGE-L. As a result, all proposals rely on NLG-oriented metrics in their original works; however, the majority lack values for chest-RRG-oriented metrics due to their novelty.

Nevertheless, NLG-oriented metrics may lack precision in clinical settings as they do not gauge syntax, semantics, or the comprehension of report meanings. Consequently, metrics tailored to the RRG task, like F_1cXb and F_1RG_{ER}, have emerged to evaluate pathologies' presence, absence, or uncertainty and their relationships within reports. These chest-RRG-oriented metrics offer a more accurate report quality assessment than NLG-oriented metrics. For instance, despite minimal differences in BLEU4 scores between Delbrouck et al. [12] and our approach in MIMIC-CXR, a significant gap is observed when measured with chest-RRG-oriented metrics. Another example of the imprecision of NLG-oriented metrics in assessing report quality can be seen in nearly identical BLEU4 scores between Miura et al. [11]'s models trained with NLL and RL in MIMIC-CXR, yet exhibiting a 12-point difference in F_1cXb. High values in NLG-oriented metrics do not reflect the quality of a report, as neither BLEU4 nor ROUGE-L measures whether the mentioned pathologies are present, absent, or uncertain.

Table 3. Comparison of our model, SwinB+BERT9k, against top state-of-the-art models for chest RRG on the MIMIC-CXR test set and Open-i dataset. The metric values are provided in (%). Bold font indicates the best result obtained for each metric.

State of the Art	**Chest-RRG-Oriented**		**NLG-Oriented**	
	F_1cXb	**F_1RG_{ER}**	**BLEU4**	**ROUGE-L**
MIMIC-CXR: NLL models				
Yang et al. [28]	-	-	11.5	28.4
Pan et al. [29]	-	-	11.2	**28.8**
Yang et al. [30]	-	-	11.1	27.4
Zhao et al. [31]	-	-	10.9	27.5
Nicolson et al. [32]	-	-	**12.7**	28.6
Liu et al. [5]	29.2	-	7.6	-
Chen et al. [9]	34.6	-	8.6	27.7
Miura et al. [11]	44.7	-	11.5	-
Chen et al. [10]	40.5	-	10.6	27.8
Delbrouck et al. [12]	44.8	20.2	10.5	25.3
Ours: NLL stage	**57.8**	**27.1**	10.3	22.8
MIMIC-CXR: RL models				
Miura et al. [11] (BERTScr+$fact_{ENT}$)	56.7	-	11.1	-
Miura et al. [11] (BERTScr+$fact_{ENTNLI}$)	56.7	-	11.4	-
Delbrouck et al. [12] (BERTScr+F_1RG_{ER})	62.2	34.7	11.4	26.5
Ours: RL stage (BERTScr+F_1RG_{ER})	**66.2**	**37.8**	**13.2**	**27.1**
Open-i: NLL models				
Miura et al. [11]	32.2	-	**12.1**	28.8
Chen et al. [10]	-	-	12.0	**29.8**
Alfarghaly et al. [26]	-	-	11.1	28.9
Donahue et al. [49]	-	-	9.9	27.8
Delbrouck et al. [12]	33.1	26.4	11.4	-
Ours: NLL stage	**46.6**	**34.5**	11.2	23.4
Open-i: RL models				
Miura et al. [11] (BERTScr+$fact_{ENT}$)	48.3	-	12.0	-
Miura et al. [11] (BERTScr+$fact_{ENTNLI}$)	47.8	-	13.1	-
Delbrouck et al. [12] (BERTScr+F_1RG_{ER})	49.1	41.2	13.9	32.7
Ours: RL stage (BERTScr+F_1RG_{ER})	**54.7**	**45.6**	**15.1**	**34.3**

4. Discussion

This study presents an efficient workflow for training VED models explicitly designed for chest RRG. This workflow demonstrates remarkable performance in chest-RRG-oriented metrics. One pivotal finding lies in our analysis of the proposed architectures, where we show that utilizing the SwinB encoder yields superior radiograph representations. Notably, the word-based approach proves more competitive in the decoder component than its subword-based counterpart. This finding contrasts with strategies commonly employed in current large language models (LLMs). However, this discrepancy can be attributed to the inherent variability and scarcity of data in the chest RRG task. Furthermore, our ablation study underscores the critical role of TA in effectively leveraging data for this task, demonstrating substantial improvements in metrics over strategies like image augmentation or HNM.

A prevalent issue in RRG proposals is the high repetition and monotony frequently encountered in reports generated by other state-of-the-art approaches. Typically, models tend to converge to a local minimum, yielding nearly identical standardized reports across most patients. This phenomenon arises as deep learning algorithms settle at a midpoint where the generated reports broadly apply to most cases. These reports adhere to a template naming common pathologies in a fixed order and syntactic structure, irrespective of the patient. Consequently, this pattern results in reports of dubious quality, often overlooking less common pathologies, thereby undermining the reliability of diagnostic models.

Given the high repetition and monotony in the generated reports of state-of-the-art models, we propose an analysis of report diversity using N-grams in the MIMIC-CXR test set. This analysis allows for measuring diversity by examining the number of unique N-grams present in generated reports, as illustrated in Figure 2. Specifically, we compare reports generated by our model with those produced by the approach presented in Delbrouck et al. [12], acknowledged for being the most competitive proposal in the field, as outlined in Table 3. The graph illustrates a notably greater diversity in our generated reports compared to those of the Delbrouck et al. [12] model. Moreover, our model achieves a diversity difference averaging nearly three times their model.

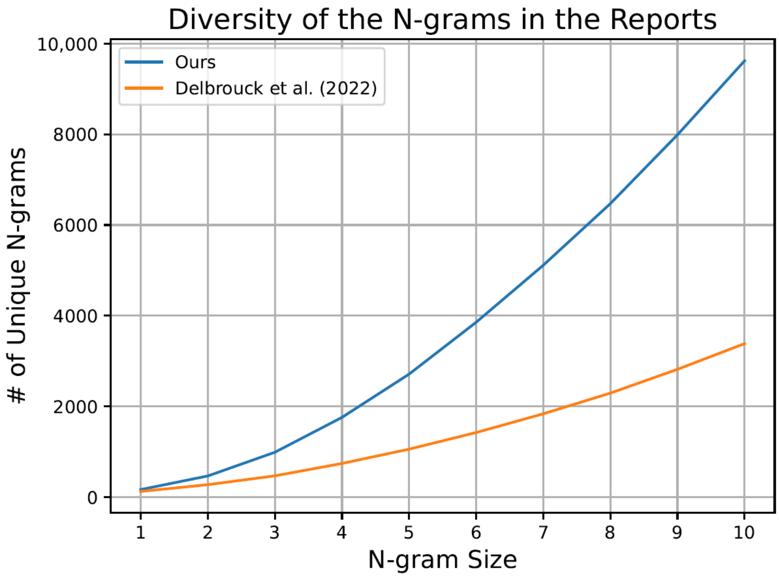
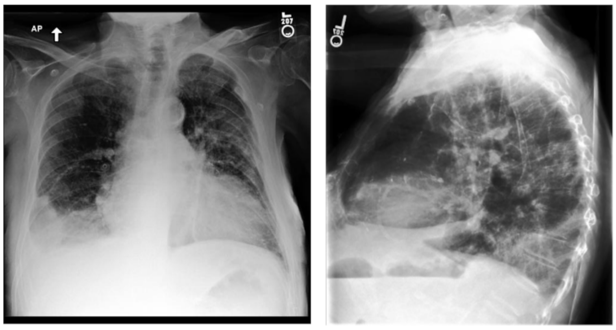
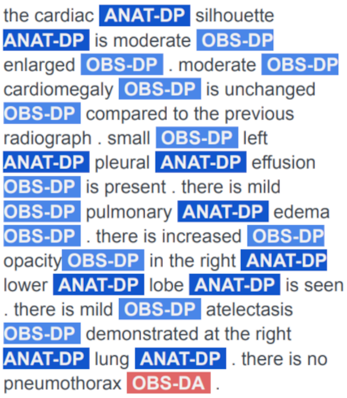
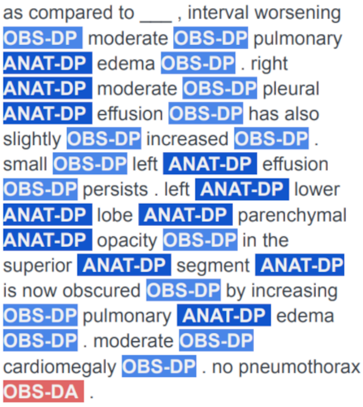
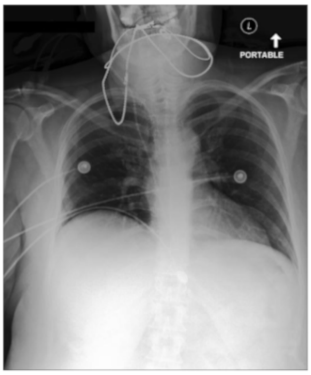
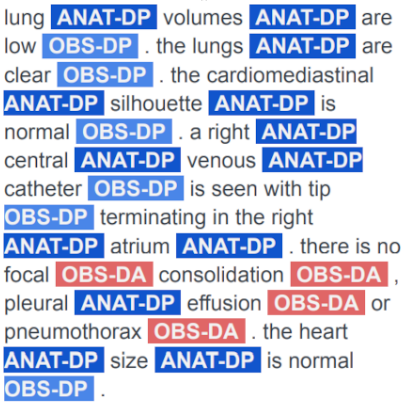
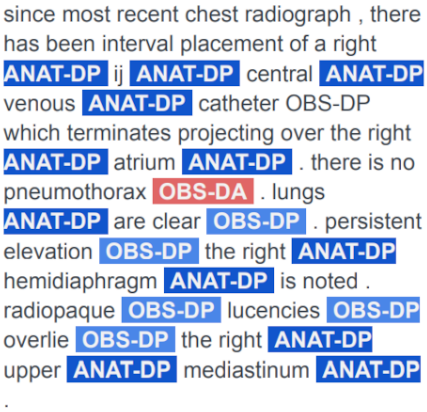
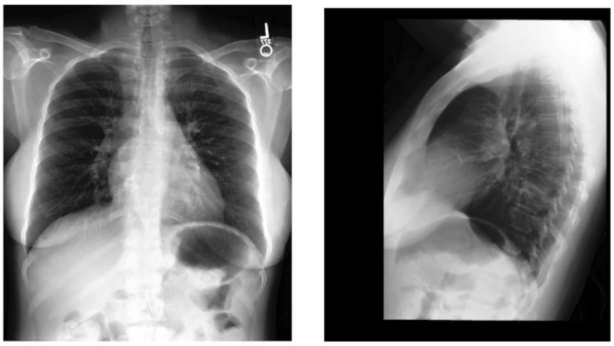
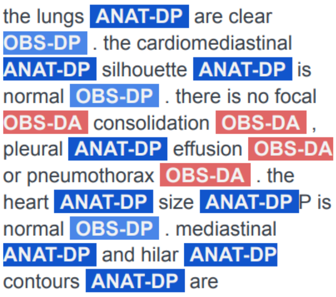
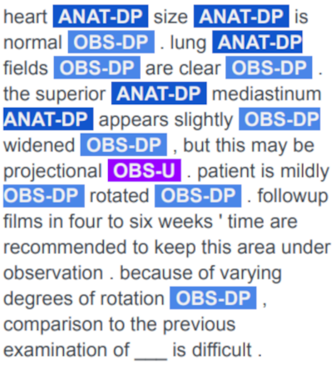


Figure 2. The graph showcases the diversity of N-grams within the generated reports using the MIMIC-CXR test set, encompassing the proposal by Delbrouck et al. [12] and our approach.

Consequently, the TA strategy substantially augments the volume of reports for training and significantly boosts the performance of VED models, as demonstrated by chest-RRG-oriented metrics. An additional qualitative advantage of our TA-based approach lies in its capacity to generate less repetitive and more diverse reports compared to prior methods. We illustrate this with examples of reports from SwinB+BERT9k and their corresponding RadGraph entities in Table 4. ANAT-DP denotes anatomical body parts mentioned in the report, while OBS signifies observations associated with radiology images, categorized as definitely present (OBS-DP), definitely absent (OBS-DA), or uncertain (OBS-U). Despite improving report quality with techniques like TA, specific repetitions and monotony persist. This will need to be addressed using new metrics and strategies in the future.

Table 4. Comparison of reports generated by our SwinB+BERT9k model trained with RL and TA against reference reports on the MIMIC-CXR test set. The first and third cases involve a patient whose study contains two images, while the second involves a single image.

Images	Hypothesis Report	Reference Report
Input images = 2	F_1RG_{ER} 63.6 %	
	 the cardiac ANAT-DP silhouette ANAT-DP is moderate OBS-DP enlarged OBS-DP . moderate OBS-DP cardiomegaly OBS-DP is unchanged OBS-DP compared to the previous radiograph . small OBS-DP left ANAT-DP pleural ANAT-DP effusion OBS-DP is present . there is mild OBS-DP pulmonary ANAT-DP edema OBS-DP . there is increased OBS-DP opacity OBS-DP in the right ANAT-DP lower ANAT-DP lobe ANAT-DP is seen . there is mild OBS-DP atelectasis OBS-DP demonstrated at the right ANAT-DP lung ANAT-DP . there is no pneumothorax OBS-DA .	 as compared to ___ , interval worsening OBS-DP moderate OBS-DP pulmonary ANAT-DP edema OBS-DP . right ANAT-DP moderate OBS-DP pleural ANAT-DP effusion OBS-DP has also slightly OBS-DP increased OBS-DP . small OBS-DP left ANAT-DP effusion OBS-DP persists . left ANAT-DP lower ANAT-DP lobe ANAT-DP parenchymal ANAT-DP opacity OBS-DP in the superior ANAT-DP segment ANAT-DP is now obscured OBS-DP by increasing OBS-DP pulmonary ANAT-DP edema OBS-DP . moderate OBS-DP cardiomegaly OBS-DP . no pneumothorax OBS-DA .
Input images = 1	F_1RG_{ER} 43.2 %	
	 lung ANAT-DP volumes ANAT-DP are low OBS-DP . the lungs ANAT-DP are clear OBS-DP . the cardiomediastinal ANAT-DP silhouette ANAT-DP is normal OBS-DP . a right ANAT-DP central ANAT-DP venous ANAT-DP catheter OBS-DP is seen with tip OBS-DP terminating in the right ANAT-DP atrium ANAT-DP . there is no focal OBS-DA consolidation OBS-DA , pleural ANAT-DP effusion OBS-DA or pneumothorax OBS-DA . the heart ANAT-DP size ANAT-DP is normal OBS-DP .	 since most recent chest radiograph , there has been interval placement of a right ANAT-DP ij ANAT-DP central ANAT-DP venous ANAT-DP catheter OBS-DP which terminates projecting over the right ANAT-DP atrium ANAT-DP . there is no pneumothorax OBS-DA . lungs ANAT-DP are clear OBS-DP . persistent elevation OBS-DP the right ANAT-DP hemidiaphragm ANAT-DP is noted . radiopaque OBS-DP lucencies OBS-DP overlie OBS-DP the right ANAT-DP upper ANAT-DP mediastinum ANAT-DP .
Input images = 2	F_1RG_{ER} 27.6 %	
	 the lungs ANAT-DP are clear OBS-DP . the cardiomediastinal ANAT-DP silhouette ANAT-DP is normal OBS-DP . there is no focal OBS-DA consolidation OBS-DA , pleural ANAT-DP effusion OBS-DA or pneumothorax OBS-DA . the heart ANAT-DP size ANAT-DPP is normal OBS-DP . mediastinal ANAT-DP and hilar ANAT-DP contours ANAT-DP are	 heart ANAT-DP size ANAT-DP is normal OBS-DP . lung ANAT-DP fields OBS-DP are clear OBS-DP . the superior ANAT-DP mediastinum ANAT-DP appears slightly OBS-DP widened OBS-DP , but this may be projectional OBS-U . patient is mildly OBS-DP rotated OBS-DP . followup films in four to six weeks ' time are recommended to keep this area under observation . because of varying degrees of rotation OBS-DP , comparison to the previous examination of ___ is difficult .

In addition to the inherent complexity of this task, focusing on medical text generation necessitates the analysis and proposal of additional chest-RRG-oriented metrics based on NLP models. Exploring novel methods to assess the reference report against an estimated one can facilitate a more comprehensive evaluation of state-of-the-art models. Our study demonstrates the utility of F_1cXb in verifying whether generated reports discuss pathologies present in the reference reports. Furthermore, F_1RG_{ER} ensures that the present, absent, and unknown pathologies are correctly related in the reports. Considering these findings, these two metrics are essential for model comparison. Although these metrics effectively and efficiently enhance RL training, models frequently produce reports with incomplete sentences. This deficiency is evident in numerous state-of-the-art proposals, and despite the efficacy of TA, it still occurs occasionally. It is evidenced at the end of the third report in Table 4. This phenomenon arises from models learning that specific syntactic structures can benefit chest-RRG-oriented metrics like F_1RG_{ER}, even if they do not constitute syntactically correct sentences. Moving forward, we advocate for further research into novel metrics using alternative deep learning algorithms, such as LLMs like GPT-4 [50] or LLaMA [51].

Several aspects of the proposed approach for real-world RRG scenarios merit attention, including multilanguage support, standardization of reports, and improvements in picture archiving and communication systems (PACS). Since the decoder model primarily operates with English words, it encounters specific difficulties when used in other languages. Thus, it would be necessary to remove the last layer of the decoder model and adjust it based on the desired language; however, the rest of the model remains language-independent. Another challenge lies in the lack of standardization in report writing across different institutions, leading to variations in reporting guidelines. Consequently, RRG models can assist in integrating new reporting guidelines to standardize and unify databases for RRG. Furthermore, our approach can enhance PACS, which are commonly used in healthcare for securely storing and transmitting electronic images and clinical reports. Collaborating with expert radiologists, we can integrate RRG models to streamline processes within PACS, such as preparing automatic draft reports that radiologists can interactively correct. This approach can improve diagnostic efficiency and expert assistance times while enhancing the standardization and homogenization of reports.

The strengths and weaknesses of our model have been highlighted using different metrics. However, another essential aspect to analyze is the integration of the proposed workflow into the clinical setting. Given that the system relies on a VED deep learning model, its deployment only requires a computer with a GPU with at least 12 GB of memory. The model specializes in chest radiographs from different position views, such as anteroposterior, lateral, posteroanterior, and lateral decubitus. Therefore, it can integrate and analyze a wide variety of chest radiographs.

The scope of this work is focused on chest radiographs. However, our approach is not limited solely to its application in chest radiographs. The proposed approach is flexible and independent, allowing it to be applied to other radiological scenarios. Therefore, our approach can be seamlessly adapted for radiographs of different body parts. The only aspect requiring modification is the reward optimization during RL training. Since F_1RG_{ER} primarily specializes in chest-related pathologies, this metric should be replaced with another relevant to the radiological scenario where our approach is intended for application. The remaining components, such as the encoder, decoder, TA, and decoding strategies, are independent of the radiological scenario. Despite the flexibility and independence of the approach, the major limitation for application to other radiological scenarios depends on the availability of training data. Acquiring large datasets to train VED models is crucial for improving the quality of reports and enhancing their applicability to different scenarios.

5. Conclusions

Chest RRG poses a challenge due to the limited data availability and the diverse nature of medical reports and variations in pathology expressions. Four transformer-based models were analyzed, highlighting the encoder as the key component, with SwinB being the best choice. This transformer encoder has not been previously explored in RRG. Regarding the decoder, a word-based approach trains faster and achieves more competitive results than a subword-based one.

Techniques like image data augmentation, hard negative mining, and TA were introduced. TA was shown to be an effective method to improve generalization and the quality of the generated reports regarding variability, quality, factual correctness, and completeness, which yields a model that outperforms the state of the art. Moreover, this approach paves the way for new TA approaches based on augmentation with more complex NLP techniques, such as LLM models.

Additionally, NLG-oriented metrics may not be optimal for measuring report quality, as they compare sentences and words without considering semantic meaning or alternative expressions of the same diagnosis. Thus, metrics based on NLP models specialized in chest radiology reports like F_1cXb and F_1RG_{ER} appear to be crucial in evaluating report quality.

Author Contributions: Investigation, D.P.; methodology, D.P.; software, D.P.; supervision, A.A. and R.P.; writing—original draft, D.P.; writing—review and editing, A.A. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: Work was partially supported by the Generalitat Valenciana under the predoctoral grant CIACIF/2022/289, with the support of valgrAI—Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and cofunded by the European Union.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao, J.; Lv, Y.; Zeng, Q.; Wan, L. Online Policy Learning-Based Output-Feedback Optimal Control of Continuous-Time Systems. *IEEE Trans. Circuits Syst. II Express Briefs* **2024**, *71*, 652–656. [CrossRef]
2. Gale, W.; Oakden-Rayner, L.; Carneiro, G.; Bradley, A.P.; Palmer, L.J. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv* **2018**, arXiv:1806.00340.
3. Li, Y.; Liang, X.; Hu, Z.; Xing, E.P. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, BC, Canada, 3–8 December 2018; pp. 1530–1540.
4. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Summers, R.M. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9049–9058.
5. Liu, G.; Hsu, T.M.H.; McDermott, M.; Boag, W.; Weng, W.H.; Szolovits, P.; Ghassemi, M. Clinically Accurate Chest X-ray Report Generation. In Proceedings of the 4th Machine Learning for Healthcare Conference, Ann Arbor, MI, USA, 9–10 August 2019; pp. 249–269.
6. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13753–13762.
7. Liu, F.; You, C.; Wu, X.; Ge, S.; wang, S.; Sun, X. Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 16266–16279.
8. Windsor, R.; Jamaludin, A.; Kadir, T.; Zisserman, A. Vision-Language Modelling For Radiological Imaging and Reports In The Low Data Regime. In Proceedings of the Medical Imaging with Deep Learning, Nashville, TN, USA, 10 July 2023; pp. 53–73.
9. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating Radiology Reports via Memory-driven Transformer. *arXiv* **2020**, arXiv:2010.16056.
10. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal Memory Networks for Radiology Report Generation. *arXiv* **2022**, arXiv:2204.13258.
11. Miura, Y.; Zhang, Y.; Tsai, E.B.; Langlotz, C.P.; Jurafsky, D. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. *arXiv* **2021**, arXiv:2010.10042.
12. Delbrouck, J.B.; Chambon, P.; Bluethgen, C.; Tsai, E.; Almusa, O.; Langlotz, C.P. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. *arXiv* **2022**, arXiv:2210.12186.
13. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675.
14. Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.Q.; Duong, D.N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.P.; Ng, A.Y.; et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv* **2021**, arXiv:2106.14463.
15. Schlegl, T.; Waldstein, S.M.; Vogl, W.D.; Schmidt-Erfurth, U.; Langs, G. Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks. In Proceedings of the Information Processing in Medical Imaging, Isle of Skye, UK, 28 June–3 July 2015; pp. 437–448.
16. Shin, H.C.; Roberts, K.; Lu, L.; Demner-Fushman, D.; Yao, J.; Summers, R.M. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2497–2506.

17. Moradi, M.; Madani, A.; Gur, Y.; Guo, Y.; Syeda-Mahmood, T. Bimodal Network Architectures for Automatic Generation of Image Annotation from Text. In Proceedings of the Medical Image Computing and Computer Assisted Intervention, Granada, Spain, 16 September 2018; pp. 449–456.
18. Rubin, J.; Sanghavi, D.; Zhao, C.; Lee, K.; Qadir, A.; Xu-Wilson, M. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. *arXiv* **2018**, arXiv:1804.07839.
19. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2577–2586.
20. Monshi, M.M.A.; Poon, J.; Chung, V. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* **2020**, *106*, 101878. [CrossRef]
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25, pp. 1097–1105.
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Liao, Y.; Liu, H.; Spasić, I. Deep learning approaches to automatic radiology report generation: A systematic review. *Inform. Med. Unlocked* **2023**, *39*, 101273. [CrossRef]
26. Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; Fahmy, A. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* **2021**, *24*, 100557. [CrossRef]
27. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
28. Yang, S.; Wu, X.; Ge, S.; Zhou, S.K.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. [CrossRef] [PubMed]
29. Pan, Y.; Liu, L.J.; Yang, X.B.; Peng, W.; Huang, Q.S. Chest radiology report generation based on cross-modal multi-scale feature fusion. *J. Radiat. Res. Appl. Sci.* **2024**, *17*, 100823. [CrossRef]
30. Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S.K.; Xiao, L. Radiology report generation with a learned knowledge base and multi-modal alignment. *Med. Image Anal.* **2023**, *86*, 102798. [CrossRef] [PubMed]
31. Zhao, G.; Zhao, Z.; Gong, W.; Li, F. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. *Artif. Intell. Med.* **2023**, *146*, 102714. [CrossRef] [PubMed]
32. Nicolson, A.; Dowling, J.; Koopman, B. Improving chest X-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **2023**, *144*, 102633. [CrossRef]
33. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
35. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597.
36. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
41. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461.
42. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
43. Johnson, A.E.W.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
44. Demner-Fushman, D.; Antani, S.; Simpson, M.; Thoma, G.R. Design and development of a multimodal biomedical information retrieval system. *J. Comput. Sci. Eng.* **2012**, *6*, 168–177. [CrossRef]

45. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
46. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 4–10 July 2004; pp. 74–81.
47. Zhang, Y.; Merck, D.; Tsai, E.B.; Manning, C.D.; Langlotz, C.P. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. *arXiv* **2020**, arXiv:1911.02541.
48. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv* **2020**, arXiv:2004.09167.
49. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
50. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
51. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.