Semantic-Aware Chest X-ray Report Generation with Domain-Specific Lexicon and Diversity-Controlled Retrieval

Baochang Zhang *[1,2,3], Chen Jia*[1], Shuting Liu[1], Heribert Schunkert[2,4], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures,
Technical University of Munich, Munich, Germany
baochang.zhang@tum.de
[2] German Heart Center Munich, Munich, Germany
[3] Munich Center for Machine Learning, Munich, Germany
[4] German Centre for Cardiovascular Research,
Munich Heart Alliance, Munich, Germany

Abstract. Image-to-text radiology report generation aims to produce comprehensive diagnostic reports by leveraging both X-ray images and historical textual data. Existing retrieval-based methods focus on maximizing similarity scores, leading to redundant content and limited diversity in generated reports. Additionally, they lack sensitivity to medical domain-specific information, failing to emphasize critical anatomical structures and disease characteristics essential for accurate diagnosis. To address these limitations, we propose a novel retrieval-augmented framework that integrates exemplar radiology reports with X-ray images to enhance report generation. First, we introduce a diversity-controlled retrieval strategy to improve information diversity and reduce redundancy, ensuring broader clinical knowledge coverage. Second, we develop a comprehensive medical lexicon covering chest anatomy, diseases, radiological descriptors, treatments, and related concepts. This lexicon is integrated into a weighted cross-entropy loss function to improve the model's sensitivity to critical medical terms. Third, we introduce a sentence-level semantic loss to enhance clinical semantic accuracy. Evaluated on the MIMIC-CXR dataset, our method achieves superior performance on clinical consistency metrics and competitive results on linguistic quality metrics, demonstrating its effectiveness in enhancing report accuracy and clinical relevance. The code is publicly available at github.com/DrLS.

Keywords: Medical Report Generation · Diversity-Controlled Retrieval · Chest-Specific Lexicon.

[1] * The two authors contributed equally to this paper.

1 Introduction

Radiology reports play a critical role in conveying accurate medical information and facilitating communication between healthcare providers and patients. However, the process is highly specialized, time-consuming, and prone to inconsistencies due to variations in radiologists' expertise. Automatic Medical Report Generation (AMRG) aims to automate the creation of structured reports from medical images, reducing the radiologist's workload while ensuring consistency and comprehensive diagnostic descriptions. The rapid development of deep learning, has greatly advanced AMRG. Early systems relied on template-based approaches [20,10], which lacked flexibility and scalability. Accurate lesion description in AMRG (including severity, localization and size) is challenging due to visual deviations, driving research in lesion classification [16], detection [13], and segmentation [17]. The shift to deep learning models, combining CNNs for image feature extraction and RNNs for text generation [22,4,18], enabled end-to-end report generation.

As AMRG models evolve [19], memory-driven architectures have significantly improved the coherence and accuracy of generated reports. Approaches such as the memory-driven transformer [3] and cross-modal memory networks [2] have advanced the alignment of visual and textual data, strengthening the semantic connection between medical images and radiology reports. To address the complexity of medical language, knowledge-driven methods have also been introduced. These techniques incorporate general and case-specific knowledge, leading to more clinically relevant reports [24]. The use of learned knowledge bases and multi-modal alignment has further refined report generation, ensuring that the generated reports maintain clinical significance [23]. Recently, incorporating structural entity extraction and patient-specific indications into report generation ensures that the content aligns more closely with clinical findings [12]. Meanwhile, retrieval-based deep learning methods [5,11,24,23,12] have gained increasing attention due to their ability to leverage historical reports for guidance. However, these models primarily focus on maximizing retrieval similarity, often leading to repetitive content and limited diversity in retrieved reports. This issue is particularly problematic when dealing with subtle variations in imaging findings or rare disease cases. Moreover, most AMRG models rely on standard cross-entropy loss [23,14,12], which treats all words equally, neglecting the importance of domain-specific terms critical for clinical decision-making. Addressing these challenges requires a more adaptive retrieval mechanism and loss optimization that prioritizes medically significant terms, ensuring more diverse, precise, and clinically meaningful report generation.

In this paper, we propose a novel method, **DrLS**, which integrates **D**iversity-controlled **r**etrieval strategy, domain-specific **L**exicon and sentence-level **S**emantic loss to enhance radiology report generation. Our contributions are as follows. (1) We propose a diversity-controlled retrieval strategy that reduces redundancy and enhances information coverage by retrieving complementary content rather than repetitive information, ensuring a broader clinical context in generated reports. (2) We introduce a medical lexicon-weighted cross-entropy loss that explicitly

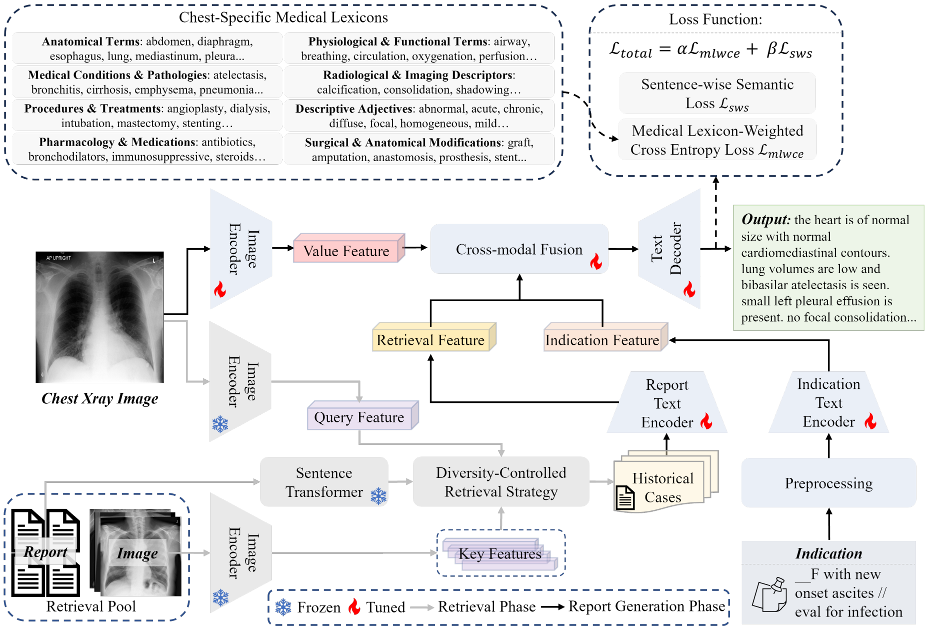

Chest-Specific Medical Lexicons
Anatomical Terms: abdomen, diaphragm, esophagus, lung, mediastinum, pleura...
Physiological & Functional Terms: airway, breathing, circulation, oxygenation, perfusion…
Medical Conditions & Pathologies: atelectasis, bronchitis, cirrhosis, emphysema, pneumonia...
Radiological & Imaging Descriptors: calcification, consolidation, shadowing…
Procedures & Treatments: angioplasty, dialysis, intubation, mastectomy, stenting…
Descriptive Adjectives: abnormal, acute, chronic, diffuse, focal, homogeneous, mild…
Pharmacology & Medications: antibiotics, bronchodilators, immunosuppressive, steroids…
Surgical & Anatomical Modifications: graft, amputation, anastomosis, prosthesis, stent...
Loss Function:
$\mathcal{L}_{total} = \alpha\mathcal{L}_{mlwce} + \beta\mathcal{L}_{sws}$
Sentence-wise Semantic Loss $\mathcal{L}_{sws}$
Medical Lexicon-Weighted Cross Entropy Loss $\mathcal{L}_{mlwce}$
AP UPRIGHT
L
Chest Xray Image
Image Encoder
Value Feature
Cross-modal Fusion
Text Decoder
Output: the heart is of normal size with normal cardiomediastinal contours. lung volumes are low and bibasilar atelectasis is seen. small left pleural effusion is present. no focal consolidation...
Retrieval Feature
Indication Feature
Image Encoder
Query Feature
Report Text Encoder
Indication Text Encoder
Sentence Transformer
Diversity-Controlled Retrieval Strategy
Historical Cases
Preprocessing
Report
Image
Retrieval Pool
Image Encoder
Key Features
Indication
___F with new onset ascites // eval for infection
❄ Frozen 🔥 Tuned → Retrieval Phase → Report Generation Phase

Fig. 1. The overview of our proposed DrLS framework.

prioritizes critical medical chest-specific terms, ensuring the model is sensitive to chest-specific clinical content while maintaining linguistic fluency. (3)We also introduce a sentence-level semantic loss that helps the model focus on the meaning of sentences, rather than merely on grammar or superficial word matching, ensuring the generated reports are more useful, reliable, and aligned with clinical expectations. (4)We evaluate our method on the MIMIC-CXR dataset, demonstrating its superiority over existing methods and its effectiveness in enhancing both the accuracy and clinical relevance of the generated radiology reports.

2 Method

An overview of the proposed method is illustrated in Fig. 1, which can be mainly divided into three parts that are introduced hereafter.

2.1 Diversity-Controlled Retrieval Strategy

Inspired by the effectiveness of Determinantal Point Processes (DPP) in optimizing subset selection [9], we propose a diversity-controlled retrieval strategy to reduce redundancy and ensure broader clinical knowledge coverage, as shown in Fig. 1. Given a set of candidate historical reports S, we define a positive semi-definite kernel matrix L, where each element is computed as $L_{ij} = q_i \cdot s_{ij} \cdot q_j$, where q_i represents the relevance score between the historical case X-ray image I_i and the current X-ray image I_q. This score is computed based on the feature vectors of I_i and I_q, extracted via the cross-modal alignment module in the pre-trained SEI method [12], and is defined as $q_i = \text{sim}(f(I_i), f(I_q))$. Here, $\text{sim}(\cdot, \cdot)$ represents the dot product similarity measure. Additionally, s_{ij} quantifies the textual similarity between the reports of historical cases R_i and R_j, which is computed using cosine similarity based on a pre-trained Sentence Transformer [15]. Then, DPP determines the selection probability of a subset $Y \subseteq S$ by computing the determinant of its corresponding submatrix:

$$P(Y) = \frac{\det(L_Y)}{\sum_{Y' \subseteq S} \det(L_{Y'})} \tag{1}$$

where $\det(L_Y)$ represents the determinant of the submatrix corresponding to the selected subset Y. The denominator sums over all possible subsets $Y' \subseteq S$, ensuring that $P(Y)$ is properly normalized as a probability distribution. To efficiently select a diverse and representative subset, we adopt a greedy optimization algorithm, which iteratively selects the report that maximizes determinant gain:

$$\arg\max_{i \in S-Y} \left[\log\det(L_{Y \cup \{i\}}) - \log\det(L_Y)\right]. \tag{2}$$

2.2 Medical Lexicon-Weighted Cross-Entropy Loss

To enhance the model's sensitivity to critical medical information, we construct a chest-specific medical lexicon and integrate it into a weighted cross-entropy loss function, emphasizing important anatomical structures and disease-related terms. The construction of the medical lexicon follows a combination of automated extraction and manual review to ensure its medical relevance and high quality. First, we use RadGraph [7] to analyze radiology reports and extract anatomical structures and observations as candidate medical terms. Then, we apply regular expressions (Regex) to extract potential medical terms while removing irrelevant short words (e.g., "cm" and numbers), followed by lowercasing and deduplication to ensure uniqueness. To further refine the high-quality medical terms, we utilize the GPT-4o-mini API for term filtering across the predefined eight categories, as shown in Fig.1, ensuring that only the most relevant anatomical structures, diseases, observations, radiological descriptors, and medical conditions are retained for radiology reports. Finally, our research team manually reviews the GPT-processed terms, removing redundant or irrelevant words to form a stable medical lexicon V_{med}.

To emphasize the importance of medical terms, we introduce a medical lexicon-weighted mechanism into the general language model used loss function (cross-entropy loss), enhancing the model's focus on critical medical information. Given a generated report token $\tilde{y}$, we define the medical lexicon-weighted cross-entropy loss $\mathcal{L}_{mlwce}$ as follows:

$$\mathcal{L}_{mlwce} = -\sum_{i=1}^{M} \omega(\tilde{y}_i) \cdot \log P(\tilde{y}_i | I, h_k, p, \{\tilde{y}_{j|j<t}\}) \tag{3}$$

$$\omega(y_{i,t}) = \begin{cases} \lambda_{med}, & \text{if } y_{i,t} \in V_{med} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where $\omega(\tilde{y}_i)$ is the weight for i_{th} predicted token $\tilde{y}_i$, and $M, I, h_k, p, \{\tilde{y}_{j|j<t}\}$ denote the maximum length of tokens generated by the text decoder, the X-ray image being processed, the set with $k = 2$ historical cases, indication prompt and the tokens generated up to time step t, respectively. Here, $\lambda_{med} > 1$ is the weighting factor for medical terms, which amplifies the loss associated with key medical words, ensuring the model prioritizes them during training.

2.3 Semantic-Level Loss for Clinical Coherence

In the task of radiology report generation, traditional cross-entropy loss primarily focuses on word-level matching, often neglecting overall semantic consistency. This limitation can lead to generated text that deviates from medical facts at the sentence level. To address this issue, we propose a semantic-level loss that enhances the global semantic representation of generated reports, ensuring that the model captures the deep semantic information inherent in medical texts. Our semantic loss is applied to the log-softmax probability distribution output by the text decoder, rather than being applied to the generated text. Specifically, given the log-softmax probability distribution $log_\sigma(P(\tilde{y}|I, h_k, p))$, we design a differentiable temperature-scaled soft-argmax function to obtain the approximate argmax result, which is formulated as,

$$\mathcal{S}_\tau(\tilde{y}) = \sigma\left(\frac{\exp(log_\sigma(P(\tilde{y}|I, h_k, p)))}{\tau}\right) \tag{5}$$

where $\tau = 0.001$ is a temperature hyperparameter that controls the smoothness of the softmax function, thus enabling the differentiable approximation of the argmax result. To obtain sentence-level semantic representations, a fine-tuned SciBERT [1] on radiology reports is used, especially the encoder f_{en} and word embedding layer f_{emb}. The word embeddings of the soft prediction $\mathcal{S}_\tau(\tilde{y})$ is obtained by multiplying it with the embedding matrix W_{emb} of the word embedding layer, $\mathcal{S}_\tau(\tilde{y}) \cdot W_e$. For the word embeddings of ground truth text can be easily obtained, $f_{emb}(y)$. Then, the proposed sentence-wise semantic loss $\mathcal{L}_{sws}$ is defined as,

$$\mathcal{L}_{sws} = 1 - \cos(f_{en}(\mathcal{S}_\tau(\tilde{y})), f_{en}(f_{emb}(y))) \tag{6}$$

where $cos(\cdot, \cdot)$ is the cosine similarity function. and $f_{en}(*)$ is the sentence-wise embedding of the input $*$, which is obtained from the the encoder's f_{en} output by taking the first token (the [CLS] token) of the last hidden state. Finally, our total loss function combines medical lexicon-weighted cross-entropy loss $\mathcal{L}_{mlwce}$ and this sentence-wise semantic loss $\mathcal{L}_{sws}$, weighted by α and β to balance word-level accuracy and semantic consistency, with the overall optimization objective formulated as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{mlwce} + \beta\mathcal{L}_{sws} \tag{7}$$

2.4 Implementation details

For network architecture, we utilize ResNet101 [6] as the image encoder, SciBERT [1] as the text encoder, and a memory-driven Transformer from R2Gen [3] as the text decoder. To reduce computational costs, we first retrieve the top 100 most similar historical cases based on image similarity and then apply our proposed diversity-controlled retrieval strategy to optimize report selection. We train our model on an NVIDIA GPU (Quadro RTX A6000) using the AdamW optimizer with an initial learning rate of 1.0×10^{-5}, a weight decay of 7.0×10^{-5}, and a batch size of 32. The loss parameters are set to $\alpha = 0.7$, $\beta = 0.3$, and $\lambda_{med} = 2$.

3 Experiments and Results

3.1 Dataset and Evaluation Metrics

We conduct experiments on MIMIC-CXR v2.0.0 [8], a large-scale public dataset of chest X-rays collected from Beth Israel Deaconess Medical Center in Boston, containing 377,110 images from 227,835 radiographic studies. We adopt the official MIMIC-CXR split, where the original data includes 368,960 images for training, 5,159 for testing, and 2,991 for validation. For consistency and fair comparison, we follow the preprocessing approach in the SEI method [12], removing samples without reports or containing anomalies, resulting in a final dataset of 269,239 images for training, 2,113 images for validation, and 3,852 images for testing.

We employ two sets of evaluation metrics: linguistic quality metrics and clinical consistency metrics, following those used in SEI work [12]. The linguistic quality metrics including BLEU-2 (BL-2), BLEU-4 (BL-4), METEOR (MTR) and ROUGE-L (R-L). Specifically, BL-2 and BL-4 assess lexical similarity by measuring n-gram overlap between the generated and reference reports. MTR provides a more comprehensive evaluation of lexical similarity by considering precision, recall, synonymy, stemming, and word order. R-L measures the longest common subsequence between the generated and reference reports, evaluating both precision and recall. The clinical consistency metrics focus on assessing factual accuracy and clinical correctness, including $F_{1,mic-5}$ CheXbert (CX5), $F_{1,mic-14}$ CheXbert (CX14), and F_1 RadGraph (RG). Meanwhile, we evaluate the generated reports of our method and comparison approaches with the ground truth report truncated at various lengths $M_{gt} \in \{60, 80, 90, 100, Cpl.\}$ to assess coherence and integrality comprehensively, where $Cpl.$ means the complete length of reference reports.

3.2 Comparison with State-of-the-art

Table 1 presents a detail comparison between our method and several mainstream radiology report generation models, including R2Gen [3], R2GenCMN [2],

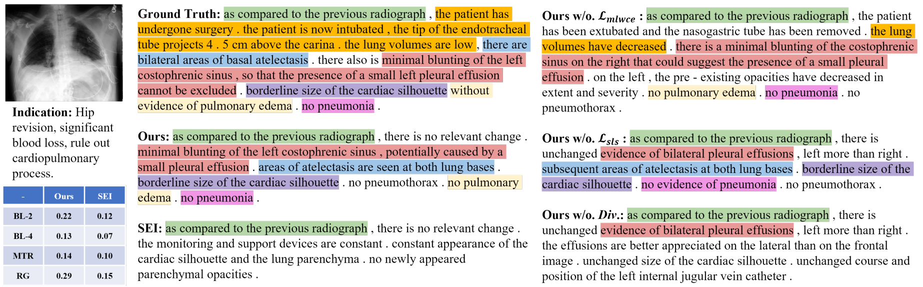
Table 1. Comparison of our method with state-of-the-art approaches on MIMIC-CXR. The best values with $\boldsymbol{M_{gt}}$ of Cpl. are highlighted in **bold**

Methods	$\boldsymbol{M_{gt}}$	**BL-2**	**BL-4**	**MTR**	**R-L**	**RG**	**CX5**	**CX14**
R2Gen [3]	100	0.218	0.103	0.137	0.264	0.207	0.340	0.340
	Cpl.	0.209	0.097	0.135	0.266	0.211	0.339	0.338
R2GenCMN [2]	100	0.218	0.106	0.142	0.278	0.220	0.461	0.278
	Cpl.	0.198	0.090	0.133	0.268	0.223	0.464	0.393
CGPT2 [14]	60	0.248	0.127	0.155	0.286	0.223	0.463	0.391
	Cpl.	0.204	0.102	0.138	0.277	0.237	0.483	0.434
M2KT [23]	80	0.237	0.111	0.137	0.274	0.204	0.477	0.352
	Cpl.	0.204	0.085	0.133	0.244	0.210	0.483	0.413
RGRG [21]	Cpl.	**0.249**	0.126	**0.168**	0.264	-	0.547	0.447
SEI [12]	60	0.268	0.148	0.167	0.301	0.236	0.509	0.445
	80	0.257	0.140	0.162	0.300	0.247	0.535	0.457
	90	0.251	0.137	0.160	0.300	0.248	0.539	0.459
	100	0.247	0.135	0.158	0.299	0.249	0.542	0.460
	Cpl.	0.238	0.128	0.154	0.296	0.249	0.545	0.460
Ours(DrLS)	60	0.265	0.151	0.164	0.309	0.250	0.511	0.451
	80	0.254	0.143	0.162	0.308	0.265	0.549	0.469
	90	0.248	0.140	0.160	0.307	0.268	0.558	0.473
	100	0.245	0.138	0.159	0.307	0.269	0.562	0.475
	Cpl.	0.240	**0.135**	0.157	**0.305**	**0.271**	**0.565**	**0.477**

Table 2. Ablation study on MIMIC-CXR. "w/o *Div.*": replacing by the similarity historical cases retrieval used in the SEI-baseline, "w/o $\mathcal{L}_{mlwce}$": degrading to standard cross entropy loss, and "w/o $\mathcal{L}_{sls}$": without the semantic loss. SEI-Baseline refers to the original SEI model.

Methods	$\boldsymbol{M_{gt}}$	**BL-2**	**BL-4**	**MTR**	**R-L**	**RG**	**CX5**	**CX14**
SEI-Baseline	100	0.247	0.135	0.158	0.299	0.249	0.542	0.460
Ours(DrLS)	100	0.245	0.138	0.159	**0.307**	**0.269**	**0.562**	**0.475**
w/o. *Div.*	100	0.245	0.136	0.158	0.303	0.263	**0.562**	0.469
w/o. $\mathcal{L}_{mlwce}$	100	0.245	0.135	0.157	0.303	0.260	0.552	0.471
w/o. $\mathcal{L}_{sws}$	100	**0.250**	**0.140**	**0.162**	**0.307**	0.267	0.554	**0.475**

CGPT2 [14], M2KT [23], RGRG [21] and SEI [12]. Compared to these methods, our model achieves the best performance on Cpl.-wise evaluation across all clinical consistency metrics, with RG of 0.271, CX5 of 0.565, and CX14 of 0.477, demonstrating its superior reliability in generating clinically accurate and factually consistent reports. Additionally, it remains strong competitiveness in linguistic quality metrics, ranking 2nd in BL-2 (0.240), 1st in BL-4 (0.135), 2rd



Indication: Hip revision, significant blood loss, rule out cardiopulmonary process.

-	Ours	SEI
BL-2	0.22	0.12
BL-4	0.13	0.07
MTR	0.14	0.10
RG	0.29	0.15

Ground Truth: as compared to the previous radiograph , the patient has undergone surgery . the patient is now intubated , the tip of the endotracheal tube projects 4 . 5 cm above the carina . the lung volumes are low , there are bilateral areas of basal atelectasis . there also is minimal blunting of the left costophrenic sinus , so that the presence of a small left pleural effusion cannot be excluded . borderline size of the cardiac silhouette without evidence of pulmonary edema . no pneumonia .

Ours: as compared to the previous radiograph , there is no relevant change . minimal blunting of the left costophrenic sinus , potentially caused by a small pleural effusion . areas of atelectasis are seen at both lung bases . borderline size of the cardiac silhouette . no pneumothorax . no pulmonary edema . no pneumonia .

SEI: as compared to the previous radiograph , there is no relevant change . the monitoring and support devices are constant . constant appearance of the cardiac silhouette and the lung parenchyma . no newly appeared parenchymal opacities .

Ours w/o. $\mathcal{L}_{mlwce}$: as compared to the previous radiograph , the patient has been extubated and the nasogastric tube has been removed . the lung volumes have decreased . there is a minimal blunting of the costophrenic sinus on the right that could suggest the presence of a small pleural effusion . on the left , the pre - existing opacities have decreased in extent and severity . no pulmonary edema . no pneumonia . no pneumothorax .

Ours w/o. $\mathcal{L}_{sls}$: as compared to the previous radiograph , there is unchanged evidence of bilateral pleural effusions , left more than right . subsequent areas of atelectasis at both lung bases . borderline size of the cardiac silhouette . no evidence of pneumonia . no pneumothorax .

Ours w/o. *Div*.: as compared to the previous radiograph , there is unchanged evidence of bilateral pleural effusions , left more than right . the effusions are better appreciated on the lateral than on the frontal image . unchanged size of the cardiac silhouette . unchanged course and position of the left internal jugular vein catheter .

Fig. 2. Generated radiology report examples. Colors highlight key observations in the ground truth, with matched segments in the generated reports shown in the same colors, illustrating model performance in capturing clinical information.

in MTR (0.157), and 1st in R-L (0.305). We further present qualitative examples of generated reports, comparing report completeness and clinical relevance. As shown in Figure 2, our method generates reports with more comprehensive clinical information than the baseline SEI method, further validating its effectiveness in radiology report generation.

3.3 Ablation Study

To assess the contribution of each component in our method, we conduct an ablation study on the MIMIC-CXR dataset, as shown in Table 2. Replacing our diversity-controlled retrieval with the similarity-based historical case retrieval used in the SEI-Baseline [12] reduces R-L and RG, highlighting its role in enhancing clinical factual completeness. Degrading our proposed medical lexicon-weighted cross-entropy loss $\mathcal{L}_{mlwce}$ to a standard cross-entropy loss results in declines across all metrics, confirming its importance in improving both lexical overlap and clinical accuracy. Meanwhile, removing the sentence-wise semantic loss $\mathcal{L}_{sws}$ slightly improves linguistic quality metrics but lowers RG and CX5, suggesting a trade-off between fluency and factual correctness. These findings demonstrate that each component contributes uniquely to report generation, with our full model achieving the best balance between linguistic quality and clinical relevance.

4 Conclusion

We propose DrLS, a novel framework for radiology report generation that integrates a diversity-controlled retrieval strategy, a medical lexicon-weighted loss, and a sentence-wise semantic loss to enhance both clinical accuracy and linguistic coherence. Our approach reduces redundancy in retrieved reports, prioritizes key medical terms through a domain-specific lexicon, and improves factual consistency by aligning sentence-level semantics. Extensive experiments on the

MIMIC-CXR dataset demonstrate that our method outperforms existing models in clinical consistency metrics while maintaining strong performance in linguistic quality. Ablation studies further confirm the effectiveness of each component. Extensive experiments on the MIMIC-CXR dataset demonstrate the effectiveness of our approach in producing clinically precise and linguistically coherent radiology reports, underscoring its potential for real-world medical applications.

Acknowledgments. The project was supported by the Bavarian State Ministry of Science and Arts within the framework of the "Digitaler Herz-OP" project under the grant number 1530/891 02 and the China Scholarship Council (File No.202004910390). We also thank BrainLab AG for their partial support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: 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). pp. 5904–5914 (2021)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
4. Dong, Y., Pan, Y., Zhang, J., Xu, W.: Learning to read chest x-ray images from 16000+ examples using cnn. In: 2017 IEEE/ACM international conference on connected health: applications, systems and engineering technologies (CHASE). pp. 51–57. IEEE (2017)
5. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health. pp. 209–219. PMLR (2021)
6. 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. pp. 770–778 (2016)
7. 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 preprint arXiv:2106.14463 (2021)
8. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
9. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)
10. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6666–6673 (2019)

11. 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. pp. 13753–13762 (2021)
12. Liu, K., Ma, Z., Kang, X., Zhong, Z., Jiao, Z., Baird, G., Bai, H., Miao, Q.: Structural entities extraction and patient indications incorporation for chest x-ray report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 433–443. Springer (2024)
13. Luo, L., Chen, H., Zhou, Y., Lin, H., Heng, P.A.: Oxnet: Deep omni-supervised thoracic disease detection from chest x-rays. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 537–548. Springer (2021)
14. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
15. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
16. Shamrat, F.J.M., Azam, S., Karim, A., Ahmed, K., Bui, F.M., De Boer, F.: High-precision multiclass classification of lung disease through customized mobilenetv2 from chest x-ray images. Computers in Biology and Medicine **155**, 106646 (2023)
17. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19. IEEE reviews in biomedical engineering **14**, 4–15 (2020)
18. 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. pp. 2497–2506 (2016)
19. Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M.: Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering (2024)
20. Tange, H.J., Hasman, A., de Vries Robbé, P.F., Schouten, H.C.: Medical narratives in electronic medical records. International journal of medical informatics **46**(1), 7–29 (1997)
21. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023)
22. 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. pp. 9049–9058 (2018)
23. 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. Medical Image Analysis **86**, 102798 (2023)
24. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical image analysis **80**, 102510 (2022)