Disease-Consistent Visual-Text Alignment and Fusion for Reliable Radiology Report Generation

Jianjun Shao and Yun Tian(✉)

School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
202421080154@mail.bnu.edu.cn, tianyun@bnu.edu.cn

Abstract. Automatic medical report generation (MRG) has significant clinical value, as it can alleviate the heavy workload of report writing for radiologists. Recent studies adopt retrieval-based strategies to inject external diagnostic knowledge into report generation. However, reliable MRG remains challenging because existing methods often retrieve diagnostically inconsistent reports and lack mechanisms for selective visual-text integration. In this work, we propose DCAF, a Disease-Consistent Visual-Text Alignment and Fusion framework for reliable radiology report generation. DCAF incorporates a Disease-Consistent Retrieval Enhancer that integrates disease priors via multi-positive contrastive learning and a disease-matching constraint to improve diagnostic consistency during retrieval. To better mimic the selective diagnostic reasoning process of radiologists, we further design an Optimal Transport-Enhanced Cross-Modal Fusion module that selectively integrates clinically relevant textual knowledge. In addition, an auxiliary disease-aware feature extraction mechanism encodes predicted diagnostic hypotheses into visual representations, enabling hypothesis-driven visual reasoning. Extensive experiments on IU X-Ray and MIMIC-CXR demonstrate state-of-the-art performance in both language generation quality and clinical efficacy.

Keywords: Report Generation · Alignment · Cross-Modal Fusion

1 Introduction

Automated medical report generation can significantly improve physician productivity and has therefore received increasing research attention. Mainstream approaches [3, 15, 23] adopt encoder-decoder architectures inspired by image captioning [4, 16, 26], enhanced by Transformer-based attention mechanisms [22]. However, unlike image captioning, radiology reports emphasize clinically significant abnormalities, which are often subtle and severely imbalanced relative to normal findings, causing models to overlook critical diagnostic cues. To mitigate this issue, recent studies [11, 21, 6, 25] attempt to inject external medical knowledge through auxiliary tasks like classification [8] or by integrating reference reports via retrieval-based algorithms[2, 24].

Despite recent progress, reliable MRG remains challenging due to three fundamental limitations. First, the diagnostic reliability of retrieved reference reports remains limited. Most retrieval-based methods [2, 8] rely on pretrained

vision-language models [18] to retrieve semantically similar reports, implicitly assuming that semantic similarity is sufficient to guarantee disease consistency. This may lead to disease-level inconsistencies in retrieved reports, thereby compromising the clinical reliability of the generated outputs. Second, existing approaches [22] heavily rely on standard multi-head cross-attention mechanisms [8, 25] for cross-modal fusion. Such mechanisms establish token-level correspondences based on local similarity and treat all retrieved textual information equally. This coarse fusion paradigm fails to reflect the selective diagnostic reasoning process of radiologists, potentially amplifying clinically irrelevant information and leading to suboptimal report generation. Third, disease semantics are rarely embedded into the visual representation space. Although some studies [8] employ disease classification as an auxiliary objective, it is usually incorporated at the language level rather than directly integrated into visual feature learning. This limits the model's capacity to capture diagnostically discriminative visual patterns.

To address these challenges, we propose DCAF, a Disease-Consistent Visual-Text Alignment and Fusion framework for reliable radiology report generation. Specifically, we introduce a Disease-Consistent Retrieval Enhancer (DRE) that incorporates disease priors into image-text representation learning and retrieval. DRE employs multi-positive contrastive learning to align images with reports sharing identical clinical findings, and a disease-matching constraint enforces diagnostic consistency during retrieval. Furthermore, we design an Optimal Transport-Enhanced Cross-Modal Fusion module to selectively integrate clinically relevant textual knowledge conditioned on the visual evidence. Finally, we introduce an auxiliary disease classification branch to explicitly encode disease semantics into visual representations, enabling disease-aware feature extraction. The main contributions of this work are summarized as follows:

1. We propose a Disease-Consistent Retrieval Enhancer to improve the diagnostic reliability of retrieved reference reports via multi-positive contrastive learning with an explicit disease-matching constraint.
2. We design an Optimal Transport-Enhanced Cross-Modal Fusion module to selectively integrate clinically relevant textual knowledge, and introduce an auxiliary classifier to extract disease-aware features, enhancing the discriminability of disease-related patterns.
3. Extensive comparisons and ablation studies demonstrate consistent improvements over state-of-the-art methods across both language generation and clinical efficacy metrics.

2 Method

We present the proposed DCAF framework. As illustrated in Fig. 1, the framework incorporates three components: Disease-Consistent Retrieval Enhancer, Optimal Transport-Enhanced Cross-Modal Fusion, and Disease Classifier.

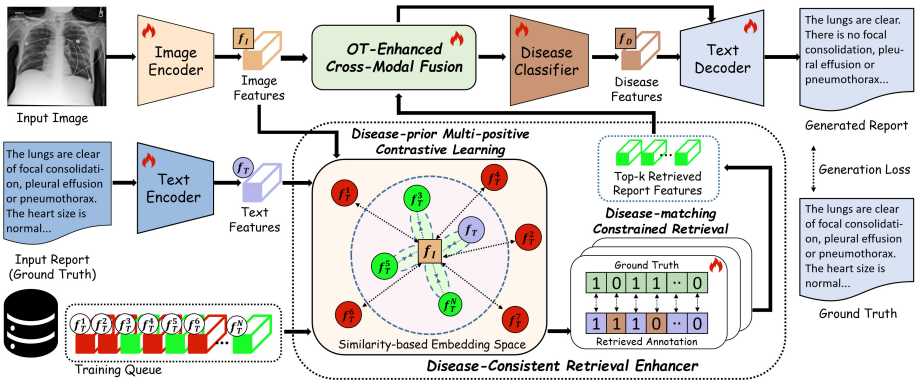

Fig. 1. Overview of DCAF, which consists of three components: 1) Disease-Consistent Retrieval Enhancer, including Multi-positive Contrastive Learning (DMCL) and Disease-matching Constrained Image-Text Retrieval (DCR); 2) Optimal Transport-Enhanced Cross-Modal Fusion; and 3) Disease Classifier.

2.1 Disease-Consistent Retrieval Enhancer

Accurate reference retrieval is critical for reliable report generation. Existing approaches typically retrieve reports based on globally pooled CLIP [18] feature similarity, which may ignore disease-level consistency. To address this issue, we propose a Disease-Consistent Retrieval Enhancer that incorporates disease priors.

Disease-prior Multi-positive Contrastive Learning First, we propose a disease-prior multi-positive contrastive learning strategy to align chest X-ray images with their corresponding diagnostic reports in a shared semantic space, such that image-text pairs with identical clinical findings are closely aligned.

Formally, given an X-ray image $I \in \mathbb{R}^{v\times h\times w}$ and its associated report $T \in \mathbb{R}^{L\times e}$, we use an image encoder and a text encoder, yielding disease-aware embeddings F_I and F_T in $\mathbb{R}^{d\times e}$ through a learnable disease-aware projection layer, where v denotes the number of views of the input image, d denotes the number of disease categories [7], L denotes the length of the report, and e is the embedding dimension. Each image-text pair (I_i, T_j) is associated with a multi-label disease annotation $y_i, y_{T_j} \in \{0,1\}^d$ extracted using CheXbert [20], where 1 signifies the confirmed presence of a specific disease, while 0 represents all other cases.

Within a training batch, reports that share identical disease labels with the input image are treated as positive samples, defined as $\mathcal{P}_i = \{T_j \mid y_{T_j} = y_i\}$, while reports with mismatched disease annotations are regarded as negative samples, defined as $\mathcal{N}_i = \{T_j \mid y_{T_j} \neq y_i\}$. Based on this formulation, we design a multi-positive contrastive objective to explicitly model disease-level consistency:

$$\mathcal{L}_{con}^{D} = -\log \frac{\sum_{j\in\mathcal{P}_i} \exp(\text{sim}(F_I, F_{T_j})/\tau)}{\sum_{j\in\mathcal{P}_i\cup\mathcal{N}_i} \exp(\text{sim}(F_I, F_{T_j})/\tau)}, \tag{1}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity and τ is a temperature parameter.

Unlike conventional contrastive learning with a single positive, our method incorporates multiple disease-consistent positives. This alleviates the sparsity issue of single-positive supervision and improves robustness to label imbalance.

Disease-matching Constrained Image-Text Retrieval Despite encouraging disease-aware alignment, retrieval solely based on embedding similarity may still introduce diagnostically inconsistent references. To address this issue, we introduce a disease-matching constraint to further improve diagnostic consistency.

Formally, given an input image embedding F_I, we first select the top-k text embeddings $\{F'_{T_1}, \ldots, F'_{T_k}\}$ from the retrieval queue based on similarity scores. Then, the constraint γ is defined as:

$$\gamma = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}_{CE}\left(y^I, y_i^T\right), \tag{2}$$

where y^I and y_i^T denote the ground-truth disease annotations of the input image and the i-th retrieved report, and k is the number of retrieved reports. $\mathcal{L}_{CE}$ represents the binary cross-entropy loss [14] computed independently for each disease category.

Minimizing this constraint encourages retrieved reports to contain disease-relevant findings similar to those in the input image.

2.2 Optimal Transport-Enhanced Cross-Modal Fusion

In clinical practice, radiologists do not treat all available textual information as equally informative. Instead, they selectively consult and weigh relevant clinical records and prior reports based on the visual findings in the image, while disregarding unrelated or redundant information.

We employ Multi-Prompt Sinkhorn Attention (MPSA) [10] for visual-text cross-attention to model this behavior. Unlike conventional multi-head cross-attention (MHCA) [22], MPSA is grounded in optimal transport (OT) theory and incorporates global contextual constraints into the attention computation, enabling globally constrained reweighting of visual-text correspondences. This allows the model to selectively emphasize clinically relevant textual knowledge, conditioned on visual evidence, while suppressing irrelevant content.

Formally, given the visual feature representation F_v extracted from the input medical image and the global textual representation F_g encoded from the most relevant retrieved clinical report, the fused features $F_c \in \mathbb{R}^{d \times e}$ are defined as:

$$F_c = \mathrm{MPSA}(QK^T)V, \tag{3}$$

where $Q = \phi_q(F_v)$, $K = \phi_k(F_g)$, and $V = \phi_v(F_g)$ denote the query, key, and value projections, respectively.

To preserve visual fidelity and stabilize training, we employ a residual connection followed by layer normalization, formulated as $F_c = f_{proj}(\mathrm{LayerNorm}(F_c + F_v))$, where f_{proj} denotes a learnable projection layer.

2.3 Disease Classifier for Disease-aware Feature Extraction

To explicitly encode disease semantics, we introduce an auxiliary disease classifier that leverages the fused features F_c to predict disease annotations. A per-disease binary Softmax is applied over each disease category. The predicted disease probabilities and corresponding classification loss are given by:

$$\hat{y} = Softmax\left(\frac{F_c \cdot \Phi^\top}{\sqrt{e}}\right), \tag{4}$$

$$\mathcal{L}_{cls} = Focal(\hat{y}, y), \tag{5}$$

where $\hat{y} \in \mathbb{R}^{d\times 2}$ denotes the predicted disease probabilities, $y \in \mathbb{R}^{d\times 2}$ represents the ground-truth disease annotations, $\Phi \in \mathbb{R}^{2\times e}$ is a learnable disease embedding matrix. The Focal term denotes the focal loss [13], which mitigates severe label imbalance in multi-label disease classification.

Based on the predicted distribution, disease-aware features $F_d \in \mathbb{R}^{d\times e}$ are extracted as $F_d = \hat{y} \cdot \Phi + F_c$, which emphasize disease-relevant visual patterns while preserving the original fused representation.

2.4 Loss Function

We construct a text decoder to synthesize a reliable report from the fused cross-modal representation F_c and the disease-aware features F_d. During training, the decoder is optimized using an autoregressive cross-entropy loss [14] $\mathcal{L}_{gen}$:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{N} \log P(T_t \mid T_{<t}, F_c, F_d), \tag{6}$$

where T_t denotes the t-th token in the ground-truth report T, $T_{<t}$ represents previously generated tokens, and N is the report length.

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{con}^{D} + \lambda_{match}\gamma + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{gen}\mathcal{L}_{gen}. \tag{7}$$

where λ_{match}, λ_{cls}, and λ_{gen} are weighting coefficients that balance retrieval consistency, disease supervision, and generation objectives.

3 Experiment

Datasets and Evaluation Metrics We evaluate our model on two widely used benchmarks: MIMIC-CXR [9] and IU X-Ray [5]. The MIMIC-CXR dataset, released by the Beth Israel Deaconess Medical Center, contains a large-scale collection of chest X-ray images paired with corresponding reports. Following previous works [2, 8], we split the dataset into 270,790 training samples, 2,130 validation samples, and 3,858 test samples for fair comparison. The IU X-Ray dataset from Indiana University includes 7,470 frontal and lateral chest X-ray

images and 3,955 associated reports. Due to the limited positive samples for certain diseases in the test set, we follow prior work [8] by evaluating the model trained on MIMIC-CXR directly on the entire IU X-Ray dataset.

We evaluate model performance in two aspects: natural language generation (NLG) and clinical efficacy (CE). For NLG, we measure report quality using BLEU [17], METEOR [1], and ROUGE-L [12]. For CE, we use CheXbert [20] to label the generated reports and compare them with ground-truth labels across 14 disease categories, assessing precision, recall, and the F1 score.

Table 1. Comparison with state-of-the-art methods on the MIMIC-CXR and IU X-Ray datasets. BL-1, BL-2, BL-4 denote BLEU scores; MTR denotes METEOR; R-L denotes ROUGE-L; Pre. denotes Precision; The best results are highlighted in bold.

Dataset	Model	NLG Metrics					CE Metrics		
		BL-1	BL-2	BL-4	MTR	R-L	Pre.	Recall	F1
MIMIC-CXR	R2Gen	0.353	0.218	0.103	0.142	0.277	0.333	0.273	0.276
	M2TR	0.378	0.232	0.107	0.145	0.272	0.240	0.428	0.308
	M2KT	0.386	0.237	0.111	0.137	0.274	0.420	0.339	0.352
	DCL	–	–	0.109	0.150	0.284	0.471	0.352	0.373
	RGRG	0.373	0.249	0.126	0.168	0.264	0.461	0.475	0.447
	KiUT	0.393	0.243	0.113	0.160	0.285	0.371	0.318	0.321
	PromptMRG	0.398	0.239	0.112	0.157	0.268	0.501	0.509	0.476
	EKAGen	0.419	0.258	0.119	0.157	0.287	0.517	0.483	0.499
	KIA	0.415	0.274	0.136	0.167	0.307	0.504	0.425	0.461
	STREAM	0.422	0.271	0.137	0.165	0.290	0.527	0.412	0.462
	Ours	**0.438**	**0.278**	**0.146**	**0.177**	**0.314**	**0.542**	**0.526**	**0.552**
IU X-Ray	R2Gen	0.289	0.155	0.052	0.128	0.243	0.151	0.145	0.145
	M2KT	0.371	0.239	0.078	0.153	0.261	0.153	0.145	0.145
	DCL	0.354	0.230	0.074	0.152	0.267	0.168	0.167	0.162
	RGRG	0.266	0.215	0.063	0.146	0.180	0.183	0.187	0.180
	PromptMRG	0.401	0.247	0.098	0.160	0.281	0.213	0.229	0.211
	Ours	**0.417**	**0.258**	**0.105**	**0.167**	**0.292**	**0.276**	**0.268**	**0.256**

Implementation details Our baseline model adopts a structure similar to that in [8], integrating a pre-trained ResNet-101 [19], a Transformer encoder, a Transformer decoder, and an auxiliary disease classification branch. The Transformer architecture consists of 8 attention heads with dimension of $e = 256$. We use the AdamW optimizer with a batch size of 32, a learning rate of 5×10^{-4}, and a weight decay of 0.01. All experiments are conducted on an NVIDIA A800 GPU (80GB).

4 Analysis

Performance Comparison To evaluate the effectiveness of our model, we compare it with a comprehensive set of state-of-the-art methods, including R2Gen

[3], M2TR [15], M2KT [23], DCL [11], RGRG [21], KiUT [6], PromptMRG [8], EKAGen [2], KIA [25], and STREAM [24]. The quantitative results are summarized in Table 1. On the MIMIC-CXR dataset, our method achieves the best performance across all NLG and CE metrics. Specifically, it attains BLEU-1 (0.438), BLEU-2 (0.278), BLEU-4 (0.146), METEOR (0.177), ROUGE-L (0.314), Precision (0.542), Recall (0.526), and an F1 score (0.552), demonstrating superior clinical consistency and report generation quality. Following PromptMRG [8], we directly evaluate models pretrained on MIMIC-CXR for the IU X-Ray dataset. As shown in Table 1, our method also achieves the highest performance across all NLG metrics and CE metrics, indicating strong cross-dataset generalization.

Table 2. Ablation study on MIMIC test set, assessing the key components: Disease-prior Multi-positive Contrastive Learning (DMCL), Disease-matching Constrained Image-Text Retrieval (DCR), Optimal Transport-Enhanced Cross-Modal Fusion (OTCF), and Disease Classifier for Disease-aware Feature Extraction (DFE).

DMCL	DCR	OTCF	DFE	NLG Metrics				CE Metrics		
				BL-1	BL-4	METEOR	R-L	Precision	Recall	F1
×	×	×	×	0.397	0.116	0.158	0.272	0.452	0.436	0.437
✓	×	×	×	0.405	0.120	0.160	0.278	0.497	0.478	0.471
✓	✓	×	×	0.416	0.126	0.164	0.286	0.522	0.508	0.524
✓	✓	✓	×	0.431	0.143	0.172	0.305	0.528	0.514	0.531
✓	✓	✓	✓	**0.438**	**0.146**	**0.177**	**0.314**	**0.542**	**0.526**	**0.552**

Ablation Studies As shown in Table 2, we conduct ablation experiments on the MIMIC-CXR test set to analyze the contribution of each component by incrementally integrating them into the BASE configuration which only includes classifier and generator. Introducing DMCL yields consistent improvements across both NLG and CE metrics, demonstrating the effectiveness of disease-aware image-text alignment. When DCR is further incorporated, additional gains are observed in BLEU and ROUGE-L, indicating that enforcing disease consistency during retrieval facilitates the selection of more clinically relevant reference reports. Adding OTCF results in further performance improvements, particularly in higher-order NLG metrics, validating its ability to selectively integrate complementary cross-modal information. Finally, integrating DFE achieves the best overall performance, highlighting the importance of explicitly modeling disease semantics for accurate and clinically consistent report generation.

Additionally, we analyzed the impact of different top-k values on report generation performance. Table 3 shows that the model achieved peak performance at $k = 18$.

Qualitative Results Fig. 2 presents a qualitative comparison between our method, the baseline, and PromptMRG. Text segments consistent with the ground truth are highlighted in colors, while inconsistent segments are shown

Table 3. Ablation study on IU X-Ray dataset, assessing the impact of different top-k values on report generation performance.

Setting	NLG Metrics						CE Metrics		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F1
$k = 6$	0.386	0.231	0.148	0.089	0.151	0.268	0.258	0.224	0.232
$k = 12$	0.406	0.251	0.162	0.101	0.162	0.288	0.261	0.249	0.247
$k = 18$	**0.417**	**0.258**	**0.166**	**0.105**	**0.167**	**0.292**	**0.276**	**0.268**	**0.256**
$k = 24$	0.415	0.254	0.164	0.103	0.165	0.290	0.273	0.263	0.251
$k = 30$	0.416	0.252	0.163	0.102	0.163	0.288	0.269	0.258	0.248

in black. Our model successfully captures the majority of key findings, including bilateral pleural effusions, pulmonary edema, and atelectasis. In contrast, PromptMRG only partially describes several critical observations, while the baseline occasionally introduces redundant or inaccurate statements. These results qualitatively demonstrate the improved clinical consistency and reliability of the proposed framework.

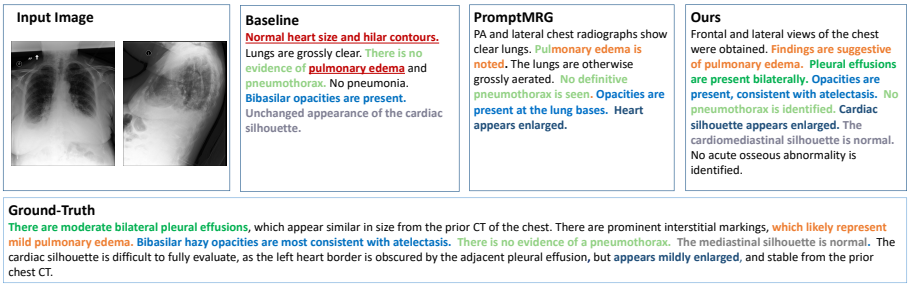

Fig. 2. Qualitative results on the MIMIC-CXR dataset are presented with visual highlights: matches with the ground truth are shown in corresponding colors, inconsistencies are marked in black, and those contradicting the ground truth are underlined in red.

5 Conclusion

We presented DCAF, a Disease-Consistent Visual-Text Alignment and Fusion framework for reliable radiology report generation. By incorporating disease-aware contrastive retrieval, an optimal transport-enhanced cross-modal fusion mechanism, and explicit disease-guided feature learning, the proposed method improves both diagnostic consistency and report generation quality. Extensive experiments on MIMIC-CXR and IU X-Ray demonstrate state-of-the-art performance across natural language generation and clinical efficacy metrics. These results underscore that DCAF provides a clinically grounded retrieval-generation paradigm for more reliable automated report generation. Future work will focus on multi-center clinical validation and real-world system integration.

Acknowledgments. This work was supported by the Beijing Natural Science Foundation Joint Fund Key Project (L251020, L256012).

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

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14194–14204 (2024)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1439–1449 (2020)
4. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
6. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19809–19818 (2023)
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 590–597 (2019)
8. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2607–2615 (2024)
9. Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R., et al.: Mimic-cxr-jpg: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
10. Kim, K., Oh, Y., Ye, J.C.: Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation. In: European Conference on Computer Vision. pp. 200–217. Springer (2024)
11. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)

13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
14. Mao, A., Mohri, M., Zhong, Y.: Cross-entropy loss functions: Theoretical analysis and applications. In: Proceedings of the International Conference on Machine Learning. pp. 23803–23828 (2023)
15. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021)
16. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10971–10980 (2020)
17. 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 of the Association for Computational Linguistics. pp. 311–318 (2002)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
19. Shafiq, M., Gu, Z.: Deep residual learning for image recognition: A survey. Applied sciences **12**(18), 8972 (2022)
20. 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 preprint arXiv:2004.09167 (2020)
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. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
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, Y., You, X., Zhang, K., Fu, Z., Wang, X., Ding, J., Sun, J., Yu, Z., Huang, Q., Han, W., et al.: Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. IEEE Transactions on Medical Imaging (2025)
25. Yin, H., Zhou, S., Wang, P., Wu, Z., Hao, Y.: Kia: Knowledge-guided implicit vision-language alignment for chest x-ray report generation. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 4096–4108 (2025)
26. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)