Learnable Retrieval Enhanced Visual-Text Alignment and Fusion for Radiology Report Generation -Supplementary-

Qin Zhou[1,2]*, Guoyan Liang[3,4]*, Xindi Li[3,4], Jingyuan Chen[3], Zhe Wang[1,2]†, Chang Yao[3,4]†, Sai Wu[3,4]†
[1]Department of Computer Science and Engineering, ECUST, China
[2]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, P. R. China
[3]Zhejiang University, Hangzhou, China
[4]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{sunniezq, wangzhe}@ecust.edu.cn, {guoyanl, 12421143, jingyuanchen, changy, wusai}@zju.edu.cn

1. Supplementary Ablation Studies

We conducted additional ablation studies on the MIMIC-CXR test set, focusing on: 1) integrating the two proposed strategies into PromptMRG[3] with CLIP, 2) analyzing key components, including hyperbolic distance and MPSA, and 3) exploring the effect of different batch size, and 4) the influence of hyperparameters.

Table 1 shows that, with the CLIP baseline, LRE alone improves average NLG and CE by 1.6%, FVTAF alone by 1.3%, and together achieve a 2.2% gain. Table 2 highlights that hyperbolic distance surpasses Euclidean distance and cosine similarity by 1.0% and 0.6%, respectively, demonstrating its strength in capturing hierarchical visual features. Table 3 reveals that replacing traditional cross-attention with MPSA provides a further 0.5% improvement, emphasizing the significance of both hyperbolic representations and MPSA. Figure 1 shows that performance declines with excessively small batches but plateaus as size increases. This is attributed to two factors: moderate batch sizes enhance input diversity through varied retrievals, while our FVTAF module's multi-source alignment design filters noise, ensuring training robustness.

As shown in Figure 2 (a) and (b), the performance remains stable across a wide range, with F1 score fluctuations within 0.8% for α and 1.4% for β. Notably, the best F1 score of 0.592 is obtained when $\alpha = 2$ and $\beta = 0.5$, which are the values we used in our experiments. As illustrated in Figure 2(a), when α deviates from 2, either too small or too large, the selection of the global reference prompt is adversely affected, resulting in lower F1 performance. Similarly, Figure 2(b) shows that the optimal performance is reached at around $\beta = 0.5$. Values of β that are too high or too low yield inferior outcomes, likely due to a multi-scale misalignment, which introduces noise and disrupts the report generation process.

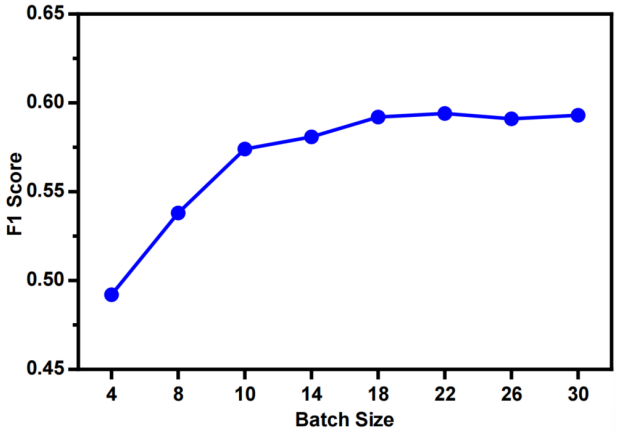

Figure 1. Effect of different batch size on model training performance.

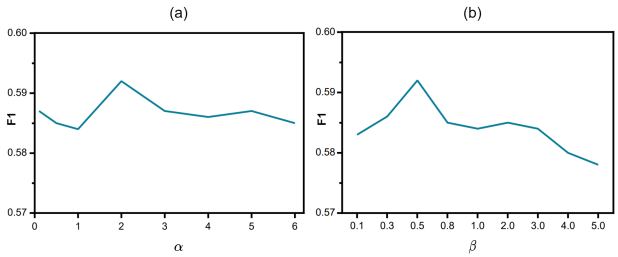

Figure 2. Analysis of the hyperparameters α (in subfigure (a)) and β (in subfigure (b)) with F1 scores on the MIMIC-CXR test set.

*These authors contributed equally.
†Corresponding Authors.

Models	NLG Metrics						CE Metrics			Avg
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Precision	Recall	F1	
Baseline (CLIP-based)	0.398	0.239	0.156	0.112	0.157	0.268	0.501	0.509	0.476	0.313
+ LRE	0.419	0.252	0.171	0.119	0.178	0.281	0.532	0.521	0.490	0.329
+ FVTAF	0.407	0.251	0.169	0.119	0.163	0.275	0.524	0.533	0.493	0.326
+ LRE & FVTAF	0.421	0.253	0.175	0.121	0.185	0.286	0.535	0.537	0.492	0.335

Table 1. Analysis on the effectiveness of each component within a CLIP-based foundation model on MIMIC-CXR test set.

Methods	NLG Metrics						CE Metrics			Avg
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Precision	Recall	F1	
EDistance	0.458	0.306	0.222	0.170	0.192	0.321	0.619	0.604	0.584	0.387
Cosine Similarity	0.461	0.315	0.232	0.181	0.197	0.333	0.618	0.603	0.583	0.391
Hyperbolic	0.465	0.318	0.235	0.182	0.199	0.336	0.628	0.613	0.592	0.397

Table 2. Comparison with different retrieval strategies on the MIMIC-CXR dataset.

2. Tackling Data Imbalance

Following the approaches [2, 3], we count the number of positive samples in the MIMIC-CXR test set and calculate the distribution of each disease, as detailed in Table 4. The results reveal a pronounced long-tailed distribution, indicating the imbalance of disease classification. For analytical clarity, we define diseases with a sample ratio exceeding 10% as head classes, and those with lower proportions as tail classes. This categorization not only facilitates a more nuanced evaluation of our model's performance but also underscores the inherent challenges associated with imbalanced data in clinical imaging datasets.

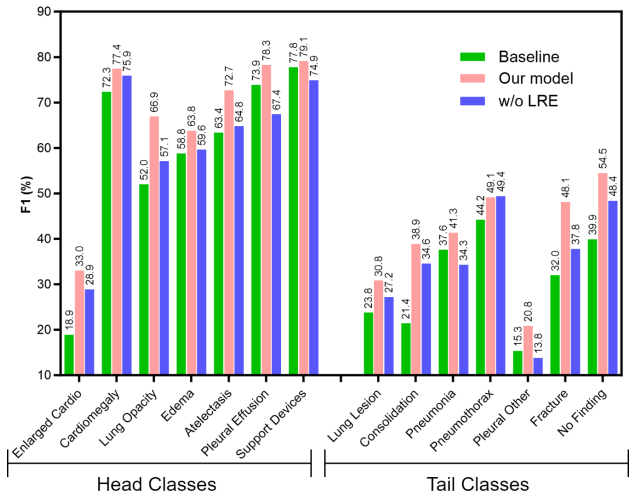

Figure 3. Comparison of baseline and our method in addressing data imbalance, evaluated using F1 scores (%).

Effectiveness of Addressing the Long-tailed Data Distribution. To evaluate the effectiveness of our method in addressing data imbalance, we categorized all diseases into head and tail groups based on sample sizes and compared the individual F1 scores between the baseline, our approach and w/o LRE module (Figure 3). Detailed head-tail grouping information is provided in the Table 3. As shown, our method consistently improves F1 scores across all disease classes, with a 9.9% average increase for tail classes and a 7.7% improvement for head classes compared to the baseline. Notably, the tail class "Fracture" achieves a remarkable 16.1% gain over the baseline. These results highlight the significant enhancements our framework brings to tail-class recognition while also delivering notable improvements for head classes. Moreover, incorporating the LRE module consistently improves performance on both head and tail classes, with average gains of 5.43% and 6.1%, respectively, compared to our model, demonstrating its effectiveness in handling data imbalance.

3. Evaluation with More Advanced Metrics

We adopt three advanced clinical evaluation metrics to comprehensively assess the effectiveness of our model, including 1/RadCliQ-v1[5], RadGraph[5], and BertScore[6]). Here we compare our model against several recent SOTA methods which are RGRG [4], MedVersa[7], and PromptMRG [3]. As shown in Table 5, our approach achieves superior performance across all metrics, outperforming the strongest model (PromptMRG) by margins of 0.11, 0.10, and 0.08, respectively. These results highlight the robustness and clinical applicability of our framework, demonstrating its capability to generate more accurate and semantically faithful reports in comparison to existing methods.

Model	NLG Metrics						CE Metrics			Avg
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Precision	Recall	F1	
Cross-attention	0.461	0.314	0.232	0.180	0.197	0.333	0.621	0.603	0.585	0.392
MSPA (Ours)	0.465	0.318	0.235	0.182	0.199	0.336	0.628	0.613	0.592	0.397

Table 3. Comparative evaluation of standard cross-attention and MPSA mechanisms on the MIMIC-CXR dataset.

	Disease Classes	Samples	Distribution
Head Classes	Enlarged Cardio	730	18.9%
	Cardiomegaly	1271	32.9%
	Lung Opacity	1392	36.1%
	Edema	563	14.6%
	Atelectasis	841	21.8%
	Pleural Effusion	1056	27.4%
	Support Devices	1345	34.9%
Tail Classes	Lung Lesion	199	5.2%
	Consolidation	176	4.6%
	Pneumonia	165	4.3%
	Pneumothorax	75	1.9%
	Pleural Other	122	3.2%
	Fracture	148	3.8%
	No Finding	323	8.4%
Total	-	3858	-

Table 4. The number of samples and their distribution ratios across disease categories in the MIMIC-CXR test set.

Model	1/RadCliQ-v1 ↑	RadGraph ↑	BertScore ↑
RGRG	0.76	0.17	0.35
MedVersa	1.10	0.27	0.45
PromptMRG	1.24	0.31	0.49
Ours	1.35	0.41	0.57

Table 5. Evaluation with advanced clinic scores on the MIMIC-CXR test set.

4. Supplementary Instruction for Evaluation on GPT-Series Multi-Modal LLMs

We use a consistent prompt for GPT-series multi-modal LLMs: "[You are helpful assistant of a radiologist. Your task is help the radiologist to draft the professional radiology report.] + [Radiology image] + [The image above is an X-ray a patients. Write a professional report on it. Answer in one paragraph, and only include the finding part.]". However, the generated output often contains extraneous information, making the evaluation unfair. To ensure consistency, we performed a two-step data cleaning process: (1) Extract only the "Findings" section from the reports and consolidate it into an individual report. (2) Remove numbers, line breaks, and other unnecessary elements, limiting the text length to 200 characters. The final evaluation results, as presented in the main paper, are reported using the same evaluation criteria as previous studies [1, 3].

References

[1] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 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*, pages 14194–14204, 2024. 3

[2] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022. 2

[3] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2607–2615, 2024. 1, 2, 3

[4] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 2

[5] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. 2

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 2

[7] Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation, 2025. 2