

DART: Disease-aware Image-Text Alignment and Self-correcting Re-alignment for Trustworthy Radiology Report Generation

Supplementary Material

A. Contrastive Loss

The contrastive loss is based on the CLIP loss [31], which maximizes the cosine similarity between paired image-text features (positive pairs, i.e., an image and its corresponding report) while minimizing the similarity between unpaired image-text features. The contrastive loss \mathcal{L}_{con} can be expressed as:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2} \left(\log \frac{e^{(\text{sim}(\mathbf{f}_I, \mathbf{f}_T)/\tau)}{\sum_{j=1}^q e^{(\text{sim}(\mathbf{f}_I, \mathbf{f}_T^j)/\tau)} + \log \frac{e^{(\text{sim}(\mathbf{f}_I, \mathbf{f}_T)/\tau)}{\sum_{j=1}^q e^{(\text{sim}(\mathbf{f}_I^j, \mathbf{f}_T)/\tau)} \right), \quad (9)$$

where τ is a learnable temperature parameter, \mathbf{f}_I and \mathbf{f}_T are image and text features from the input image and its corresponding report, \mathbf{f}_I^j and \mathbf{f}_T^j are the j^{th} image and text features stored in the training queue, q is the number of features in the queue, and sim represents the cosine similarity between two features. The cosine similarity between features from the input image and its corresponding report is defined as:

$$\text{sim}(\mathbf{f}_I, \mathbf{f}_T) = \frac{\mathbf{f}_I \cdot \mathbf{f}_T}{|\mathbf{f}_I| \cdot |\mathbf{f}_T|}. \quad (10)$$

B. Generation Loss

We employ a cross-entropy loss, denoted as \mathcal{L}_{gen} , to train the text generator for synthesizing accurate and trustworthy radiology reports. This loss minimizes the discrepancy between the generated report \hat{T} and the ground-truth report T , which consists of l tokens $T = \{T_1, T_2, \dots, T_l\}$. At each time step t , the model predicts the probability of the next token T_t conditioned on all previous tokens T_1, T_2, \dots, T_{t-1} . The generation loss can be defined as:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^l \log P(T_t | T_1, \dots, T_{t-1}, \mathbf{f}_D, \mathbf{f}_T, \mathbf{f}_T^1, \dots, \mathbf{f}_T^k), \quad (11)$$

where T_t is the t^{th} token in the ground-truth report T , T_1, \dots, T_{t-1} represent all preceding tokens, \mathbf{f}_D represents the disease-relevant features, \mathbf{f}_T denotes the text features, $\mathbf{f}_T^1, \dots, \mathbf{f}_T^k$ are the retrieved text features, and l is the length of the ground-truth report.

C. Qualitative Analysis

Fig. 4 presents an additional qualitative analysis of generated reports of three cases from the MIMIC-CXR dataset, including generated report from our proposed framework without self-correction (“w/o Self-Correction”). We refine the generated report of “w/o Self-Correction” by GPT-4 [1] (“Correction by GPT-4”), and our proposed framework with self-correction (“Ours”). We also show the ground-truth report and the Top-3 retrieved reports from image-to-text retrieval.

Details for Correction by GPT-4 We evaluate the refinement of generated reports using GPT-4 [1]. The goal is to assess whether large language models (LLMs) can effectively improve the quality of the generated reports by addressing omissions and enhancing coherence. We provide GPT-4 with the generated report, retrieved texts, and the input image, using the following structured prompt:

[the input image] Retrieved Patient’s Text Top-1: [the retrieved text (top-1)]. ... Retrieved Patient’s Text Top-k: [the retrieved text (top-k)]. If the generated report is [the generated report], correct the generated report.

Here, the prompt includes the input image, the top- k retrieved texts from image-to-text retrieval, which provide contextual information relevant to the input image, and the generated report from our proposed framework without self-correction (“w/o Self-Correction”).

Case 1 The “w/o Self-Correction” report provides a basic assessment, accurately identifying key findings such as “lungs” and “atelectasis at the left base.” However, it omits details regarding “pulmonary vasculature,” “pleural effusion” and “pneumothorax,” which are critical for specific analysis. On the other hand, “Correction by GPT-4” introduces additional observations, such as “hyperinflated, consistent with COPD” and “mild biapical scarring,” which are not consistent with the ground-truth.

In contrast, “Ours” generates a report that aligns with the ground-truth and accurately captures key findings. It not only confirms the absence of “pleural effusion” and “pneumothorax” but also identifies detail observations such as “opacities in the left lung base likely reflect atelectasis” and “pulmonary vasculature is normal,” which are consistent with the ground-truth. Additionally, “Ours” accurately captures the description of “mediastinal and hilar contours are normal,” demonstrating its ability to comprehensively

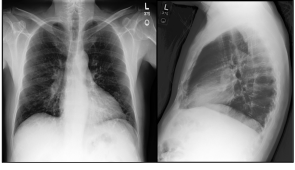
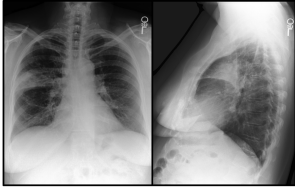
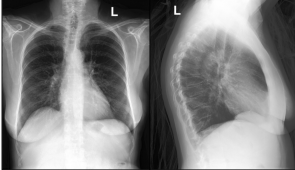
Case 1	Image	w/o Self-Correction	Correction by GPT-4	Ours
		The lungs are well expanded and clear. No focal consolidation is identified. There is minimal atelectasis at the left base. The cardiomeastinal silhouette is normal in size. Degenerative changes are noted in the thoracic spine.	The lungs are hyperinflated, consistent with COPD. No focal consolidation is identified. There is mild atelectasis at the left base, and mild biapical scarring is present. The cardiomegaly is stable, and the cardiomeastinal silhouette is normal in size. Degenerative changes are noted in the thoracic spine.	Mediastinal and hilar contours are normal. There is no focal consolidation. Opacities in the left lung base likely reflect atelectasis. Pulmonary vasculature is normal. No pleural effusion or pneumothorax is present. There are mild degenerative changes in the imaged thoracic spine.
	Ground-Truth Report	Top-1 Retrieved Text	Top-2 Retrieved Text	Top-3 Retrieved Text
	Cardiac silhouette size is mildly enlarged. Mediastinal and hilar contours are within normal limits. Pulmonary vasculature is normal. Subsegmental atelectasis is noted in the left lung base. Lungs are otherwise clear without focal consolidation. No pleural effusion or pneumothorax is present. Mild degenerative changes are noted in the thoracic spine.	Mild cardiomegaly is a stable. the lungs are hyperinflated consistent with copd. There is no pneumothorax. Right lower lobe consolidation has resolved. There is atelectasis in the left lower lobe. There is mild biapical scarring. There are mild degenerative changes in the thoracic spine.	Heart size is normal. The aorta is tortuous. Mediastinal and hilar contours are otherwise unremarkable. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities. Mild degenerative changes are noted in the mid thoracic spine.	No focal consolidation is seen. There is minimal basilar atelectasis. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.
Case 2	Image	w/o Self-Correction	Correction by GPT-4	Ours
		There is no pneumothorax or pleural effusion. There is a new opacity in the right lower lung likely representing atelectasis or consolidation concerning for pneumonia. The cardiomeastinal silhouette is within normal limits.	There is an area of opacity in the right lower lung, likely representing consolidation. The differential diagnosis includes atelectasis or consolidation concerning for pneumonia. There is no evidence of pneumothorax or pleural effusion. The cardiomeastinal silhouette is within normal limits.	Heart is top normal in size. There is a consolidation in the right lobe concerning for pneumonia. There is no pleural effusion or pneumothorax. Mediastinal contour is normal. No acute bony abnormalities.
	Ground-Truth Report	Top-1 Retrieved Text	Top-2 Retrieved Text	Top-3 Retrieved Text
	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. There is a consolidation in the anterior segment of the right upper lobe, consistent with pneumonia. Elsewhere, the lungs appear clear. There are no pleural effusions or pneumothorax. Bony structures are unremarkable.	Pa and lateral chest were provided. There is an area of consolidation at the right lung base, raises concern for pneumonia. There is no pneumothorax or pleural effusion. Cardiomeastinal silhouette is stable from prior study with the heart size being top normal.	The cardiomeastinal and hilar contours are stable. There has been interval resolution of the consolidation at the right lung base. There are no new focal consolidations concerning for pneumonia. There is no pleural effusion or pneumothorax. The pulmonary vasculature is within normal limits.	The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. No displaced fracture is seen.
Case 3	Image	w/o Self-Correction	Correction by GPT-4	Ours
		The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Heart is normal in size. Mediastinal contour is unremarkable. The bones are intact.	The lungs are clear without consolidation or edema. No pleural effusion or pneumothorax is seen. The cardiomeastinal silhouette is normal. The osseous structures are unremarkable.	The lungs are well expanded and clear. There is no pleural effusion, pneumothorax, or focal consolidation. The cardiomeastinal silhouette is normal in size. The hila are within normal limits. No acute osseous abnormalities identified.
	Ground-Truth Report	Top-1 Retrieved Text	Top-2 Retrieved Text	Top-3 Retrieved Text
	The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Heart size is normal. There is no vertebral compression fracture.	The lungs are clear without consolidation or edema. There is no pleural effusion or pneumothorax. The cardiomeastinal silhouette is normal. The osseous structures are unremarkable.	In comparison with the prior study, there is no change or evidence of acute cardiopulmonary disease. No pneumonia, vascular congestion, or pleural effusion.	Pa and lateral chest radiographs provided. Lungs are well expanded. There is no focal consolidation, pleural effusion or pneumothorax. The cardiomeastinal silhouette is normal and unchanged from the previous exam. The bones are intact.

Figure 4. An additional qualitative analysis of reports for three samples from the MIMIC-CXR dataset is presented. The top row of each sample displays an image set from two different views alongside a generated report from our proposed framework without the self-correction module (“w/o Self-Correction”). We further attempted to refine the generated report of “w/o Self-Correction” using GPT-4 [1] (“Correction by GPT-4”) to compare it with the generated report from our proposed framework with self-correction (“Ours”). The bottom row shows the ground-truth report and the Top-3 retrieved texts from image-to-text retrieval. Key findings are highlighted in different colors for clarity.

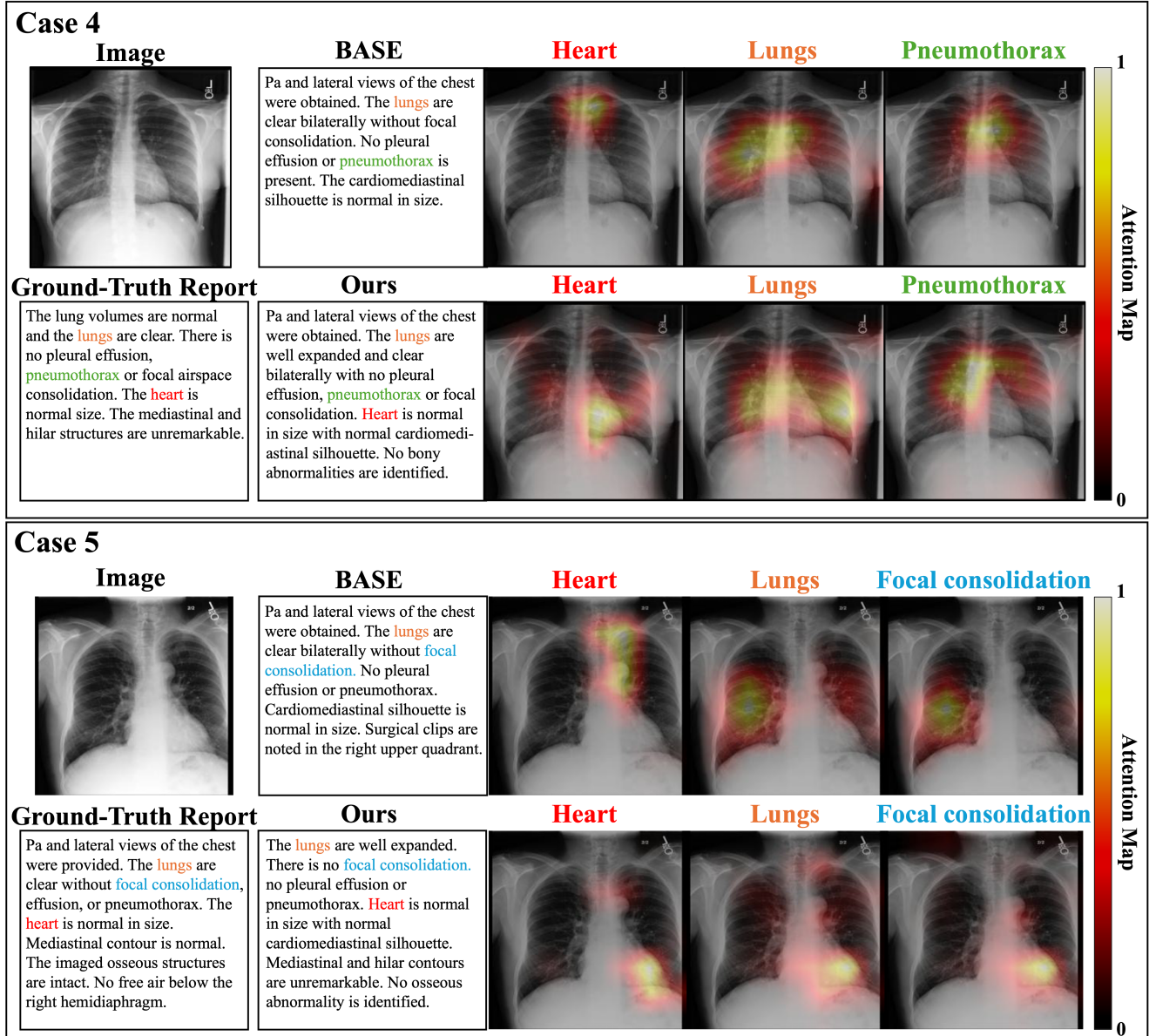


Figure 5. Visualizations of the generated reports and attention maps from the baseline model (BASE) and our proposed framework (Ours) on two samples from the MIMIC-CXR dataset. The attention maps, visualized using Grad-CAM [33], illustrate the regions that BASE and Ours focuses on according to keywords such as “heart,” “lung,” “pneumothorax,” and “focal consolidation,” with each keyword highlighted in different colors.

address key disease-relevant findings, further enhancing its alignment with the ground-truth.

In both “w/o Self-Correction” and “Ours,” the Top-3 retrieved reports provide additional contextual information and contain key findings aligned with the ground-truth report, such as “Degenerative changes” and “atelectasis.” This also demonstrates that our proposed framework effectively leverages the retrieved reports similar to the ground-

truth.

Case 2 The “w/o Self-Correction” report identifies essential findings such as the absence of “pneumothorax and pleural effusion.” However, it does not comprehensively address “mediastinal contour” or “bony structures.” Similarly, “Correction by GPT-4” refines the phrasing of findings, such as describing the opacity as “likely representing consolidation.” However, it produces redundancy and does

not explicitly describe some key findings, such as “mediastinal contour and the “bony structures.”

In contrast, “Ours” generates a report that aligns with the ground-truth and accurately captures the patient’s condition. It not only identifies the absence of “pleural effusion and pneumothorax,” but also describes the “mediastinal contour” as normal and uniquely includes a statement about the absence of acute “bony abnormalities,” aligning with the ground-truth, such as “bony structures are unremarkable.”

In both “w/o Self-Correction” and “Ours,” the Top-3 retrieved reports provide additional contextual information and contain key findings aligned with the ground-truth report, such as “consolidation,” “pneumothorax,” “pleural effusion,” and “pneumonia.” This also demonstrates that our proposed framework effectively leverages the retrieved reports, which are similar to the ground-truth.

Case 3 “w/o Self-Correction” successfully captures key findings from the ground-truth, such as “pleural effusion,” “pneumothorax,” and “consolidation.” However, both “Correction by GPT-4” and “Ours” generate the phrase “cardiomediastinal silhouette” instead of “heart.” Similarly, while the retrieved texts effectively capture key findings from the ground-truth report, such as “pleural effusion” and “pneumothorax,” they include “cardiomediastinal silhouette” instead of “heart.” The term “cardiomediastinal silhouette” can be used as an indirect indicator for assessing “heart size.” Since the retrieved texts do not include the direct keyword “heart,” self-correction mechanisms, both “Correction by GPT-4” and “Ours,” generate an indirect term instead.

This case highlights the importance of designing self-correction mechanisms to prioritize the retrieval of reports that explicitly include key findings from the ground-truth. Accurate retrieval is crucial for ensuring that generated reports align closely with disease-relevant findings. While our proposed framework demonstrates significant improvements in capturing these findings, this example underscores the need to refine the retrieval to directly align with the ground-truth report in the self-correction process.

D. Attention Visualization

Fig. 5 presents an additional attention visualization using Grad-CAM [33] to compare the BASE setting (“BASE”) and our proposed framework (“Ours”) for radiology report generation. BASE setting includes only the classification loss and generation loss. The visualization highlights the regions of focus for three critical keywords with each keyword represented in a distinct color for clarity.

Case 4 Both models successfully generate the keywords “lungs” and “pneumothorax,” aligning with the ground-truth report. However, the baseline model misses “heart,” while our proposed model accurately captures it. This difference is reflected in the attention maps: our proposed

model focuses on the actual heart region, as well as “lungs” and “pneumothorax,” whereas the baseline model fails to attend to the heart region. These results demonstrate the effectiveness of our proposed model in capturing disease-related findings.

Case 5 Both “BASE” and “Ours” successfully generate the keywords “lungs” and “focal consolidation,” aligning with the ground-truth report. However, the attention maps again highlight notable differences. Similar to Case 4, the “BASE” model attends predominantly to regions associated with the “lungs” but fails to focus on key areas related to the “heart.” Additionally, its attention for “focal consolidation” is similar with the regions of “lungs.”

For “Ours,” the attention maps exhibit strong focus on the “heart,” demonstrating the ability of our proposed framework to identify and prioritize critical regions for this keyword. However, for “lungs” and “focal consolidation,” the attention maps show some focus on irrelevant regions. Despite this limitation, our proposed framework successfully generates the keywords “lungs” and “focal consolidation,” which are clinically accurate and align with the ground-truth report. This highlights the inherent difficulty of extracting disease-relevant features directly from X-ray images. It also highlights the effectiveness of our proposed framework compared to “BASE,” particularly in leveraging retrieved reports and self-correction mechanisms to supplement and guide the report generation process, thereby compensating for potential inconsistencies with image features.

E. Ablation Study on IU X-ray

We extend our ablation study to the IU X-ray dataset to evaluate the incremental impact of each component in our proposed framework: contrastive loss (CL), image-to-text retrieval (I2T), disease-matching constraint (DM), and self-correction (SC). The results are summarized in Table 3, showing performance improvements as these components are progressively added to the BASE setting, which includes only the classification loss and generation loss.

Starting from the BASE setting, which achieves BLEU-4 of 0.124 and ROUGE-L of 0.326, the addition of contrastive learning (CL) in setting (a) leads to modest improvements in BLEU-4 (0.137) and ROUGE-L (0.355). This indicates that aligning image and text embeddings through contrastive learning enhances feature representation, which aids the downstream generation task.

Adding image-to-text retrieval (I2T) in setting (b) significantly boosts performance across all metrics, with BLEU-4 increasing to 0.174 and ROUGE-L to 0.358. This demonstrates the value of retrieving disease-relevant reports, which provide additional contextual information for accurate report generation.

In setting (c), the inclusion of the disease-matching constraint (DM) further improves performance, with BLEU-4

Dataset	Setting	CL	I2T	DM	SC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	RG-L	METEOR
IU X-ray	BASE	-	-	-	-	0.421	0.271	0.183	0.124	0.326	0.169
	(a)	✓	-	-	-	0.427	0.282	0.195	0.137	0.355	0.169
	(b)	✓	✓	-	-	0.464	0.320	0.230	0.174	0.358	0.185
	(c)	✓	✓	✓	-	0.472	0.328	0.240	0.182	0.386	0.201
	(d)	✓	✓	✓	✓	0.486	0.348	0.265	0.208	0.411	0.205

Table 4. An ablation study of our proposed framework on the IU X-ray dataset, assessing the impact of key components: contrastive loss (CL), image-to-text retrieval (I2T), disease-matching constraint (DM), and self-correction (SC). A “✓” indicates the presence of each component, while “-” denotes its absence. The BASE setting involves training only with the classification loss and the generation loss.

reaching 0.182 and ROUGE-L increasing to 0.386. The disease-matching constraint ensures that the retrieved reports align more closely with the disease-relevant findings of the input images, resulting in more accurate and clinically coherent generated reports.

Finally, adding self-correction (SC) in setting (d) achieves the best results, with BLEU-4 improving to 0.208 and ROUGE-L reaching 0.411. This substantial improvement highlights the effectiveness of the self-correction module in refining the generated reports. By re-aligning the generated reports with the input image features in the embedding space, the self-correction module reduces discrepancies and enhances the accuracy and coherence of the generated reports.

This ablation study on the IU X-ray dataset demonstrates the consistent effectiveness of each component in our proposed framework. In other words, this study validates the importance of integrating contrastive learning, disease-aware retrieval, disease-matching, and self-correction to achieve state-of-the-art performance in radiology report generation.

F. Effect of Retrieved Texts

Our proposed framework retrieves similar texts based on input images to generate accurate reports. Fig. 6 evaluates the effect of retrieved texts, ranging from $k = 0$ (without retrieval) to $k = 5$, on the BLEU-4 performance. It demonstrates that retrieving texts ($k = 1, 2, \dots, 5$) enhances the BLEU-4 score compared to the performance without retrieval ($k = 0$).

In detail, the BLEU-4 score for $k = 0$ (without retrieval) is 0.113, which is significantly lower than the BLEU-4 scores achieved when retrieval is employed. This underscores the importance of retrieval in our proposed framework. The retrieved texts provide critical disease-relevant findings that enhance the alignment between the generated reports and the ground-truth findings, thereby improving performance for report generation.

The BLEU-4 score gradually increases as k increases from 1 to 3, suggesting that retrieving more texts provides additional useful context for generating accurate radiology

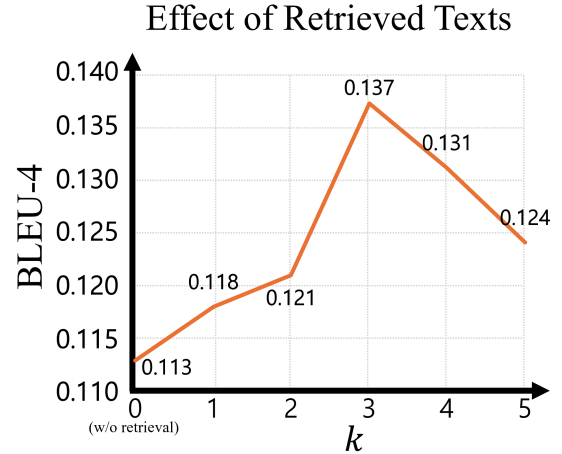


Figure 6. We evaluate the effect of the number of retrieved texts (k) on BLEU-4 performance for the MIMIC-CXR dataset in our proposed framework.

reports. However, when k exceeds 3, a decline in performance is observed. Our possible explanation is that the additional retrieved texts beyond $k = 3$ may include less relevant information, which could dilute the effectiveness of disease-relevant findings.

In summary, this analysis highlights the importance of the retrieval process in providing relevant textual information and demonstrates its crucial role in generating accurate and comprehensive radiology reports.