ClinAlign: Clinical Workflow Aligned Memory Retrieval for Radiology Report Generation

Lihong Qiao[1], **Shiyi Gao**[1], **Yucheng Shu**[1*], **Bin Xiao**[1] and **Weisheng Li**[1]
[1]School of Artificial Intelligence, Chongqing University of Posts and Telecommunications
{qiaolh, shuyc, xiaobin, liws}@cqupt.edu.cn, shiyigao27@gmail.com

Abstract

Automated radiology report generation aims to create clear and clinically correct diagnostic reports from medical images. Existing retrieval enhancement methods primarily focus on reusing textual knowledge, neglecting the crucial role of local visual pattern memory in clinical diagnosis. Furthermore, cross-modal retrieval lacking explicit clinical semantic constraints can easily introduce irrelevant pathological information, thereby reducing the clinical effectiveness of the generated reports. To address these challenges, we propose **ClinAlign**—a memory-based retrieval framework aligned with clinical workflow, drawing inspiration from clinical diagnostic workflows. Visually, we construct a disease-aware visual memory bank and enhance local patch representations through proposed Memory-based Patch Pattern Augmentation (**MPPA**), thereby improving the perception and discrimination of pathological regions. On the textual side, we construct a disease-aware textual memory bank and introduce Classification-Guided Prompt Augmentation (**CGPA**), where disease state predictions are converted into structured diagnostic prompts to provide explicit semantic guidance for textual memory retrieval. Extensive experiments on two medical report generation benchmarks, MIMIC-CXR and IU X-Ray, demonstrate the effectiveness and practical value of our proposed method.

1 Introduction

Radiology report generation aims to connect computer vision and natural language processing to automatically generate structurally standardized, professional, and clinically consistent diagnostic reports from medical images [Nooralahzadeh *et al.*, 2021; Yan and Pei, 2022; Wang *et al.*, 2023a] . This task has significant practical value for clinical decision support and reducing the workload of radiologists [Tanida *et al.*, 2023; Bu *et al.*, 2024; Tanno *et al.*, 2025] . Compared

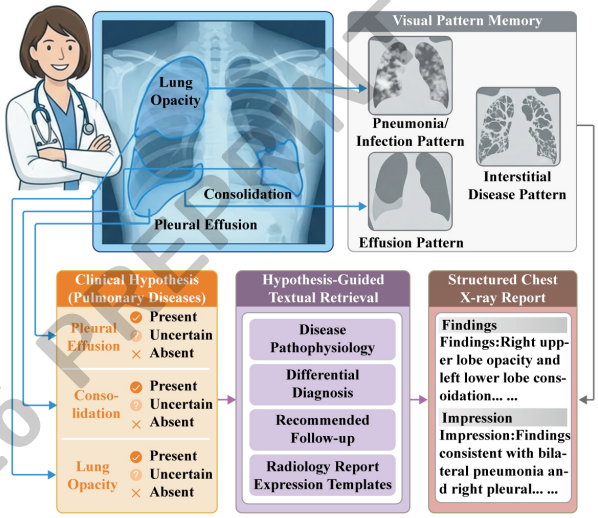


Figure 1: Clinical workflow of radiologists during image reading and report writing. Radiologists identify abnormalities, match them to visual pattern memory, and retrieve disease-state–conditioned knowledge and templates to produce structured reports.

to general image description tasks, radiology report generation is much more challenging: reports are typically lengthy, structurally complex, and contain a large amount of technical terminology. Furthermore, in radiology images, key lesions often present subtle appearance features, have low contrast characteristics, and are easily obscured by complex anatomical structures , making accurate perception and description of lesions extremely difficult [Xie *et al.*, 2023].

In recent years, various research methods have emerged in the field of radiology report generation, including encoder-decoder frameworks enhanced with knowledge graphs [Liu *et al.*, 2021; Chen *et al.*, 2022; Li *et al.*, 2023; Huang *et al.*, 2023], memory mechanisms [Jin *et al.*, 2024; Wang *et al.*, 2024c; Tao *et al.*, 2024], and large language models [Wang *et al.*, 2023b; Chen *et al.*, 2023; Liu *et al.*, 2024b]. These methods have significantly improved the quality of generated reports. However, existing methods still have shortcomings in realistically reflecting the clinical diagnostic process, especially in fine-grained lesion modeling and clinical semantic consistency. Specifically, these limitations manifest in

*Corresponding author.

the following two aspects: First, existing research mainly focuses on the reuse of textual memory, neglecting visual pattern memory at the image patch level [Jin *et al.*, 2024; Wang *et al.*, 2024c; Tao *et al.*, 2024]. In the complex realities of clinical diagnosis, radiologists rely heavily on visual memory templates to conduct in-depth and meticulous analyses of lesions' morphology, texture, and local characteristics. However, the problem is that existing models lack explicit modeling of lesion regions at the image patch level, which makes it extremely difficult for them to accurately perceive and clearly distinguish lesion regions.

Second, current retrieval enhancement models have significant shortcomings when it comes to offering clinical semantic guidance. Usually, these approaches directly match global image features with textual memory [Jin *et al.*, 2024; Wang *et al.*, 2024c]. This leads to an over-reliance on superficial similarity calculations and a lack of essential clinical reasoning support during the retrieval process . Such a design fails to take into account the well-recognized medical diagnostic principle of "identify first, then refine." Radiologists first identify diseases and determine their states, and then retrieve relevant knowledge. Skipping this crucial semantic step might unintentionally incorporate pathologically irrelevant information into the retrieval results, thus reducing the clinical effectiveness of the generated reports. Figure 1 offers a brief overview of the real-world clinical workflow adopted by radiologists: they first spot abnormalities, match them with visual pattern memory, infer disease states, and then retrieve clinically relevant knowledge and templates to produce structured reports.

To tackle these limitations, we introduce **ClinAlign**, a memory retrieval framework for radiology report generation that is aligned with clinical workflows. From the visual perspective, ClinAlign constructs a disease-aware visual memory bank and then enhances patch-level representations via *Memory-based Patch Pattern Augmentation* (**MPPA**), thereby improving the perception and discrimination of pathological regions. In terms of text, ClinAlign constructs a disease-aware textual memory bank and proposes *Classification-Guided Prompt Augmentation* (**CGPA**). This module translates the results of disease state prediction into structured diagnostic prompts, offering explicit semantic guidance for textual memory retrieval. This unified design ensures that both the visual perception and textual retrieval processes are in line with clinical diagnostic workflows, ultimately leading to the generation of more accurate and clinically coherent radiology reports.

We conduct extensive experiments on two medical report generation benchmarks. The results show that the proposed framework achieves competitive performance in report accuracy and clinical consistency, which effectively validates its efficacy and practical significance in automatic radiology report generation.

2 Related Work

2.1 Medical Radiology Report Generation

Radiology report generation aims to integrate computer vision and natural language processing technologies to automatically produce structured, professional reports from medical images . Compared to general image description tasks, radiology report generation is much more challenging: reports are typically lengthy, structurally complex, and laden with specialized terminology. Furthermore, in radiology images, key lesions often present subtle appearance features, have low contrast characteristics, and are easily obscured by complex anatomical structures [Xie *et al.*, 2023], making accurate perception and description of lesions extremely difficult.

To overcome these limitations, researchers have explored various approaches. Mainstream methods generally employ an encoder-decoder architecture and enhance the professionalism and completeness of reports by incorporating strategies such as knowledge graphs [Liu *et al.*, 2021; Chen *et al.*, 2022; Yang *et al.*, 2022; Huang *et al.*, 2023; Li *et al.*, 2023], contrastive learning [Yang *et al.*, 2023; Hou *et al.*, 2023; Wang *et al.*, 2024a], and memory mechanisms [Jin *et al.*, 2024; Wang *et al.*, 2024c; Tao *et al.*, 2024]. Furthermore, methods optimized based on human feedback [Qin and Song, 2022; Xiao *et al.*, 2025] and the application of large language models have also been proven to effectively improve the quality of generated reports [Chen *et al.*, 2023; Liu *et al.*, 2024b; Li *et al.*, 2024a; Wang *et al.*, 2025]. However, these methods still have shortcomings in the accurate perception of subtle lesions and semantic guidance that aligns with clinical diagnostic logic, which may reduce the clinical consistency of the generated reports.

2.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) has become an important technological paradigm in the field of natural language processing, demonstrating significant value in specialized tasks such as medical report generation [Singh *et al.*, 2025]. Its core mechanism involves retrieving relevant memory knowledge units from large-scale professional corpora to provide reliable references for the generation process, thereby ensuring the professional accuracy of the output content . Current retrieval-augmented RRG methods mainly retrieve text from past reports [Yang *et al.*, 2022; Jin *et al.*, 2024; Wang *et al.*, 2024c; Tao *et al.*, 2024]. As a result, these methods deviate from the clinical workflow in two aspects. First, they rarely model local visual pattern memory, which weakens the perception and discrimination of pathological regions. Second, they often jump straight to global image–text matching, with no clear clinical guidance. Without first identifying the disease and its state, retrieval can bring in irrelevant details and make the report less clinically consistent.

3 Method

3.1 Disease-Aware Multimodal Memory Bank

Disease-Aware Visual Memory Bank As illustrated in Figure 3, given a chest X-ray image of size, visual features are extracted using a pre-trained medical Vision Transformer (ViT). The image is divided into fixed non-overlapping 16×16 patches, which are then linearly projected into token embeddings and processed by Transformer encoder layers.The resulting representations are fed into a multi-label classification head to predict disease categories in the MIMIC-CXR

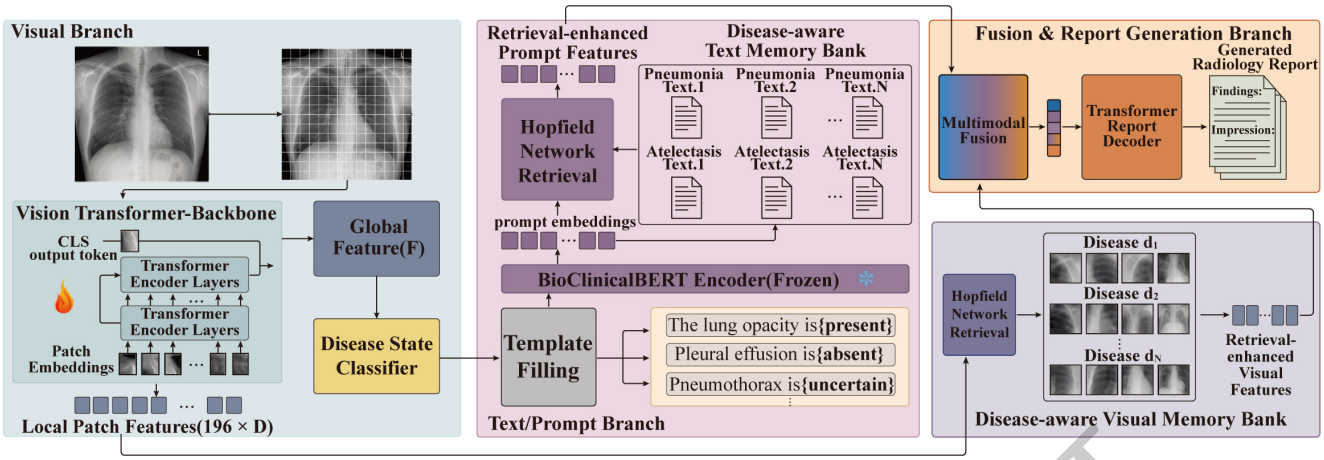


Figure 2: Overview of the ClinAlign framework. The patch-level features are enhanced via MPPA by retrieving visual pattern memory from a disease-aware visual memory bank. The text branch formulates structured prompts from the predicted disease states and applies CGPA to retrieve relevant knowledge from a disease-aware text memory bank, yielding retrieval-augmented prompt features. Finally, the retrieval-enhanced visual and textual representations are fused and decoded by a Transformer decoder to generate a structured radiology report.

dataset, with labels derived from expert-annotated radiology reports. During training, the ViT backbone remains frozen, and only the final linear classifier is fine-tuned.

Inspired by previous methods [Lin *et al.*, 2023; Wang *et al.*, 2024b], we leverage the class activation map (CAM) technique to localize disease-aware visual patches. In our work, HiResCAM is applied to highlight image regions most relevant to disease prediction. Compared with GradCAM, HiResCAM produces more spatially precise heatmaps [Draelos and Carin, 2020], enabling clearer localization of clinically significant pathological regions . Specifically, for an input X-ray image I and each of its positive disease labels, we generate a corresponding heatmap $S \in \mathbb{R}^{H \times W}$. For each image patch p_i in the heatmap S, we compute the average value v_i within the patch region. v_i is computed as follows:

$$v_i = \frac{1}{|p_i|} \sum_{(x,y) \in p_i} S(x, y), \tag{1}$$

where $|p_i|$ is the total number of pixels in patch p_i. $S(x, y)$ is the activation value at pixel (x, y) in the heatmap S .

Only when v_i exceeds a predefined threshold θ is the patch p_i retained and stored in the visual memory bank corresponding to that disease label.

$$\mathcal{V}_d = \begin{cases} \mathcal{V}_d \cup \{p_i\}, & \text{if } v_i \geq \theta \\ \mathcal{V}_d, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{V}_d$ denotes the visual memory bank for disease d. Following prior work, we set the threshold θ to 0.5.

Directly utilizing the complete visual memory bank $\mathcal{V}_d$ for training and retrieval poses significant efficiency bottlenecks and resource challenges, primarily due to its vast overall sample size. To address this issue, we compute the mean feature centroid $\mathbf{c}_d$ for all samples within $\mathcal{V}_d$:

$$\mathbf{c}_d = \frac{1}{|\mathcal{V}_d|} \sum_{\mathbf{f} \in \mathcal{V}_d} \mathbf{f} \tag{3}$$

where $|\mathcal{V}_d|$ is the total number of samples in $\mathcal{V}_d$, and $\mathbf{f}$ denotes the feature vector of an individual sample. Based on this centroid, we then select the k-nearest samples to construct a compact subset.

We apply this selection process to all disease categories, extracting a compact subset for each. The final visual memory bank $\mathcal{V}_m$ is then formed by merging all these subsets.

Disease-Aware Textual Memory Bank To leverage the rich semantic information in medical reports, we construct a textual memory bank that operates in parallel with the visual one. Specifically, we employ BioClinicalBERT [Alsentzer *et al.*, 2019] which is pre-trained on large-scale biomedical corpora — as the text encoder to extract features from all radiology reports in the training set. These textual features are then categorized according to the disease label associated with each report. For each disease d, we build its corresponding textual memory bank $\mathcal{T}_d$, formally defined as the set of textual feature vectors belonging to that disease category.

To limit the scale of the textual memory bank, we apply a parallel core sample selection strategy to all disease categories, extracting a compact and representative textual subset for each. The final textual memory bank $\mathcal{T}_m$ used for multimodal training is then constructed by merging all these disease-specific subsets.

3.2 Memory-Based Patch Pattern Augmentation

To address the lack of patch-level visual memory in existing methods. We leverage the disease-aware visual memory bank $\mathcal{V}_m$ to enhance local image features f_p extracted by the visual encoder via Hopfield-based retrieval. Following prior work [Mao *et al.*, 2025; Wang *et al.*, 2025], we adopt a modern Hopfield network as the retrieval module due to its high-capacity associative memory, robust convergence behavior, and efficient matching of high-dimensional continuous representations [Ramsauer *et al.*, 2020]. The Hopfield network performs associative retrieval between the input features and

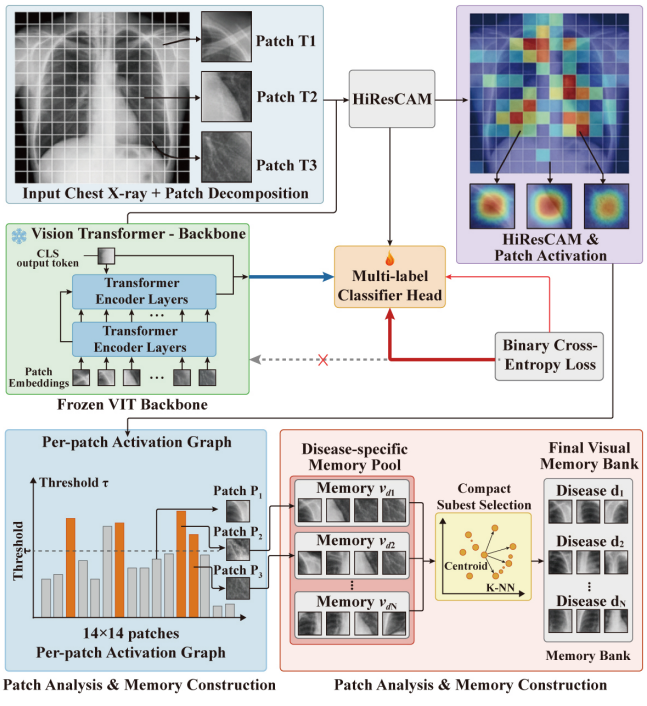


Figure 3: Pipeline for constructing the disease-aware visual memory bank.

the stored disease-aware visual prototypes, yielding an enhanced representation:

$$f_p^* = \text{Hopfield}(f_p, \mathcal{V}_m). \tag{4}$$

Specifically, the Hopfield network matches each local feature f_p against the memory patterns in $\mathcal{V}_m$ and aggregates relevant disease-aware visual information to update the feature representation. The resulting enhanced feature f_p^* explicitly incorporates disease-specific visual priors, improving model's ability to perceive and discriminate pathological visual patterns.

3.3 Classification-Guided Prompt Augmentation

Directly retrieving textual knowledge from global image features often lacks explicit clinical semantics, leading to ambiguous or irrelevant retrieval results. To provide structured semantic guidance for textual retrieval and generation, we introduce a classification-guided prompt mechanism based on disease state recognition.

We use the global image feature f_g for disease state classification, with the classification loss defined as:

$$\mathcal{L}_{class} = -\frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,m,c} \log\left(\hat{y}_{i,m,c}\right), \tag{5}$$

where M denotes the number of diseases, N the number of samples, and C the number of states per disease. $y_{i,m,c}$ and $\hat{y}_{i,m,c}$ represent the ground-truth label (one-hot) and the predicted probability, respectively.

The predicted disease states are used to fill predefined prompt templates following the pattern *"The [disease name] is [disease state] in the image."*, thereby converting structured classification outputs into natural language descriptions. The resulting prompts are encoded by BioClinicalBERT and further enhanced via Hopfield-based retrieval over the constructed disease-aware textual memory bank $\mathcal{T}_m$:

$$f_t^* = \text{Hopfield}(f_{\text{prompt}}, \mathcal{T}_m), \tag{6}$$

where f_{prompt} denotes the original prompt text feature.

Through associative retrieval from the disease-aware textual memory bank $\mathcal{T}_m$, the Hopfield network injects disease- and state-specific clinical semantics as well as typical expression patterns into the prompt representation, resulting in an enhanced textual feature that provides more informative semantic guidance for subsequent report generation.

3.4 Overall Training Objectives

Finally, the enhanced prompt feature f_t^* is fused with the retrieval-enhanced local visual feature f_p^* via a feed-forward network with SwishGLU activation to form the joint representation X. The classification results are incorporated as diagnostic prompt tokens $\mathbf{D} = \{d_1, d_2, \ldots, d_S\}$, where S denotes the fixed number of prompt tokens. Conditioned on the joint representation $\mathbf{X}$ and the diagnostic prompts $\mathbf{D}$, the text decoder F_d generates the diagnostic report $\mathbf{R}$. The decoding process is formulated as:

$$r_t = F_d(\mathbf{X}, \mathbf{D}, r_1, \ldots, r_{t-1}), \tag{7}$$

and the report generation is optimized by minimizing the language modeling loss:

$$\mathcal{L}_{LM} = -\sum_{t=1}^{T} \log p(r_t \mid \mathbf{X}, \mathbf{D}, r_1, \ldots, r_{t-1}), \tag{8}$$

where r_t denotes the token predicted at time step t, and T is the total length of the generated report.

The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{LM} + \lambda \mathcal{L}_{\text{class}}. \tag{9}$$

where λ is a balancing coefficient that controls the relative importance of the language modeling loss and the classification loss.

4 Experiments and Analysis

4.1 Experiments Settings

Datasets We evaluate our model on two public RRG benchmark datasets: IU X-Ray [Demner-Fushman *et al.*, 2015] and MIMIC-CXR [Johnson *et al.*, 2019]. IU X-Ray contains 7,470 chest X-ray image-report pairs released by Indiana University. MIMIC-CXR is a widely adopted large-scale benchmark, providing paired radiology images and diagnostic reports. For data partitioning, MIMIC-CXR uses the official split (270,790/2,130/3,858 samples). IU X-Ray splits data in a 7:2:1 ratio for reliable and fair evaluation.

Evaluation Metrics We evaluate our model with both natural language generation (NLG) metrics and clinical efficacy (CE) metrics. NLG metrics (BLEU [Papineni *et al.*, 2002], METEOR [Denkowski and Lavie, 2011], ROUGE-L [Lin, 2004]) reflect report text fluency and similarity to reference reports.

Dataset	Model	Publication	CE Metrics			NLG Metrics			
			Precision	Recall	F1	BLEU-1	BLEU-4	METEOR	ROUGE-L
MIMIC	R2Gen	EMNLP 2020	0.333	0.273	0.276	0.353	0.103	0.142	0.277
	M2TR	EMNLP 2021	0.240	0.428	0.308	0.378	0.107	0.145	0.272
	CliBert	AAAI 2022	0.397	0.435	0.415	0.383	0.106	0.144	0.275
	M2KT	MedIA 2023	0.420	0.339	0.352	0.386	0.111	–	0.274
	METrans.	CVPR 2023	0.364	0.309	0.311	0.386	0.124	0.152	0.291
	KiUT	CVPR 2023	0.371	0.318	0.321	0.393	0.113	0.160	0.285
	DCL	CVPR 2023	0.471	0.352	0.373	–	0.109	0.150	0.284
	RGRG	CVPR 2023	0.461	0.475	0.447	0.373	0.126	0.168	0.264
	HERGen	ECCV 2024	0.415	0.301	0.317	0.395	0.122	0.156	0.285
	CoFE	ECCV 2024	0.489	0.370	0.405	-	0.125	**0.176**	**0.304**
	PromptMRG	AAAI 2024	0.501	0.509	0.476	0.398	0.112	0.157	0.268
	TokenMixer	IEEE TMI 2024	-	-	-	0.409	0.124	0.158	0.288
	OaD	IEEE TMI 2024	0.364	0.382	0.372	0.412	**0.127**	0.156	0.284
	RRG-Mamba	IJCAI 2025	0.498	0.453	0.475	0.406	0.121	0.154	0.293
	Ours	–	**0.525**	**0.546**	**0.505**	**0.412**	0.121	0.163	0.278

Table 1: Results on the MIMIC-CXR dataset in terms of CE metrics and NLG metrics. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Model	Publication	NLG Metrics			
			BLEU-1	BLEU-4	METEOR	ROUGE-L
IU X-Ray	R2Gen	EMNLP 2020	0.470	0.165	0.187	0.371
	R2GenCMN	ACL 2021	0.475	0.170	0.191	0.375
	PPKED	CVPR 2021	0.483	0.168	0.187	0.376
	METrans.	CVPR 2023	0.483	0.172	0.192	0.380
	DCL	CVPR 2023	–	0.163	0.193	0.383
	CoFE	ECCV 2024	–	0.175	0.202	**0.438**
	PromptMRG	AAAI 2024	0.401	0.098	0.160	0.281
	Med-LMM	ACM MM 2024	–	0.168	**0.209**	0.381
	TokenMixer	IEEE TMI 2024	0.483	0.190	0.208	0.402
	SILC	IEEE TMI 2024	0.472	0.175	0.192	0.379
	Ours	–	**0.495**	**0.192**	0.208	0.397

Table 2: Results on the IU X-Ray dataset in terms of NLG metrics. The best results are highlighted in bold, and the second-best results are underlined.

CE metrics (precision, recall, F1-score) quantify clinical correctness by converting reports into 14 disease labels using CheXbert tool. All CE metric values reported are computed using example-based averaging.

Implementation Details The visual encoder is a medical pretrained 12-layer ViT-B/16 [Dosovitskiy, 2020], and the decoder is based on BERT-base. The training objective combines the language modeling loss and the classification loss, and the balancing coefficient λ is set to 4 based on hyperparameter ablation. We use BioClinicalBERT as the textual feature extractor. To improve retrieval efficiency, we compress both the visual and textual memory banks for each disease category by retaining the 500 samples closest to the feature centroid. This value is also determined via hyperparameter ablation. We train the model using the AdamW optimizer with a weight decay of 0.05. The initial learning rate is set to 5e-5 and decayed with a cosine schedule. We train for 15 epochs with a batch size of 16 and an input resolution of 224×224. All experiments are conducted on a single NVIDIA A40 GPU (48GB), and training takes approximately 24 hours.

4.2 Experimental Results

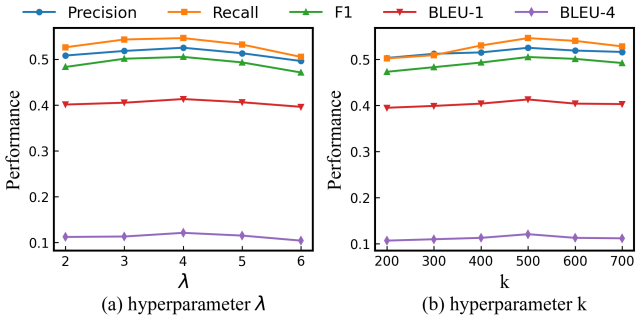
We compare our proposed method with several representative state-of-the-art radiology report generation models, including **DCL** [Li *et al.*, 2023], **KiUT** [Huang *et al.*, 2023] , **PromptMRG** [Jin *et al.*, 2024], **TokenMixer** [Yang *et al.*, 2024], **OaD** [Li *et al.*, 2024b] , **SILC** [Liu *et al.*, 2024a], and **RRG-Mamba** [Hou *et al.*, 2025]. Table 1 reports the results on the MIMIC-CXR dataset using both CE and NLG metrics, while Table 2 presents the NLG results on the IU X-Ray dataset.

Results on MIMIC-CXR As shown in Table 1, Our method achieves the best results on all three CE metrics, with 0.525 Precision, 0.546 Recall, and 0.505 F1, significantly outperforming prior models. These results indicate that our model can identify clinical entities more accurately and more completely. Specifically, compared with **DCL** (0.373 F1), our method yields an absolute improvement of 13.2% in F1; compared with **KiUT** (0.321 F1), the gain further increases to 18.4%. Even against the strongest competitor **RRG-Mamba** (0.475 F1) and **PromptMRG** (0.476 F1), our method still achieves a 3.0% absolute improvement, highlighting its superior clinical entity-level performance. Meanwhile, ClinAlign is also competitive on NLG metrics. It achieves the best BLEU-1 score (0.412) and maintains strong performance on BLEU-4, METEOR, and ROUGE-L.

In our experiments, greater emphasis is placed on clinical efficacy metrics, as these metrics truly measure whether diagnostic entities are accurate and comprehensive—which is what matters in real-world clinical settings. Text generation metrics like BLEU and ROUGE primarily assess superficial similarity between sentences. However, they often fail to reflect whether critical lesion descriptions in reports are accurate and complete. Therefore, even if some methods achieve high text scores, our approach significantly outperforms them in clinical metrics. This demonstrates its true util-

Components		Sampling Strategy		CE Metrics			NLG Metrics			
MPPA	CGPA	Random	KNN	Precision	Recall	F1	BLEU-1	BLEU-4	METEOR	ROUGE-L
✗	✗	✗	✗	0.495	0.509	0.472	0.402	0.112	0.158	0.268
✓	✗	✓	✗	0.505	0.524	0.485	0.407	0.115	0.160	0.275
✓	✗	✗	✓	0.518	0.532	0.496	0.406	0.113	0.160	0.272
✗	✓	✓	✗	0.507	0.521	0.486	0.408	0.118	0.161	0.277
✗	✓	✗	✓	0.513	0.528	0.492	0.411	0.120	0.160	0.276
✓	✓	✗	✓	**0.525**	**0.546**	**0.505**	**0.412**	**0.121**	**0.163**	**0.278**

Table 3: Ablation study of proposed components and sampling strategy on on MIMIC test set.



(a) hyperparameter λ (b) hyperparameter k

Figure 4: Performance on MIMIC test set under varying balancing coefficients λ and memory compression sizes k.

ity and reliability in real-world medical settings. In summary, our approach balances clinical correctness and language quality, significantly improving clinical efficacy while preserving the fluency of the report.

Results on IU X-Ray Since IU X-Ray does not provide standardized annotations or an established protocol for CE evaluation, prior work typically reports only NLG metrics. To guarantee an equitable comparison , we evaluate our method on IU X-Ray using NLG metrics only. As shown in Table 2, our method achieves 0.495 BLEU-1, 0.193 BLEU-4, 0.208 METEOR, and 0.397 ROUGE-L, where BLEU-1, METEOR, and ROUGE-L are the best or tied for the best. For instance, compared with **R2GenCMN** [Chen *et al.*, 2021], our method improves BLEU-1 from 0.475 to 0.495 and ROUGE-L from 0.375 to 0.397. In addition, our method outperforms strong recent baselines such as **TokenMixer** on BLEU-1 (0.495 vs 0.483) and achieves comparable performance on other NLG metrics.

Overall, our method substantially improves clinical entity-level correctness on MIMIC-CXR and achieves strong NLG performance on IU X-Ray, supporting the effectiveness of ClinAlign in improving both clinical accuracy and report generation quality.

4.3 Ablation Study

Analysis on proposed components Table 3 reports ablation results on the MIMIC test set, analyzing the contributions of MPPA and CGPA. It is important to note that without CGPA, retrieval is conducted by similarity matching between global image features and the textual memory bank. The baseline achieves CE scores of 0.495/0.509/0.472 (Precision/Recall/F1). Adding MPPA improves CE to 0.505/0.524/0.485, showing that disease-aware visual pattern memory strengthens patch-level representations and helps capture subtle lesions. Using CGPA alone boosts CE to 0.513/0.528/0.492, indicating that disease-state–based semantic cues provide clearer clinical guidance for retrieval and reduce irrelevant noise. Enabling both yields the best CE performance (0.525/0.546/0.505), demonstrating their complementarity in "local lesion perception + semantically guided retrieval." NLG metrics remain stable, suggesting improved clinical accuracy without sacrificing fluency. Overall, these results support the effectiveness of our design consistent with radiologists' diagnostic workflow.

Analysis of different sampling strategy We compare two retrieval sampling strategies: Random and KNN. Table 3 shows that KNN consistently yields better CE performance under the same component setting. For instance, without CGPA, switching from Random to KNN increases F1 from 0.485 to 0.496; with CGPA only, F1 improves from 0.486 to 0.492. This suggests that KNN retrieves more pathology-consistent references, reducing retrieval noise and improving entity-level clinical accuracy. With the full MPPA + CGPA setup, KNN achieves the best result (F1 = 0.505), showing the importance of a compact and relevant retrieved subset for clinical consistency.

4.4 Hyperparameter Analysis

We study the impact of the balancing coefficient λ and the memory compression size k on the MIMIC-CXR test set. λ balances clinical semantic alignment and report generation. As shown in Fig. 4(a), performance improves as λ increases from 2 to 4 and reaches its best at $\lambda = 4$, suggesting that stronger classification guidance helps retrieval-enhanced prompt learning. However, when λ is too large, performance drops because classification is over-weighted and language modeling is weakened, which reduces both CE and NLG scores. In contrast, a smaller λ provides too little semantic guidance. We further evaluate k under the default setting $\lambda = 4$, where k denotes the number of centroid-nearest samples retained in each disease-specific memory bank. As shown in Fig. 4(b), the best performance is achieved at $k = 500$. A smaller k reduces retrieval diversity and weakens the retrieved cues, whereas a larger k introduces redundant or less representative samples, injecting noise that harms

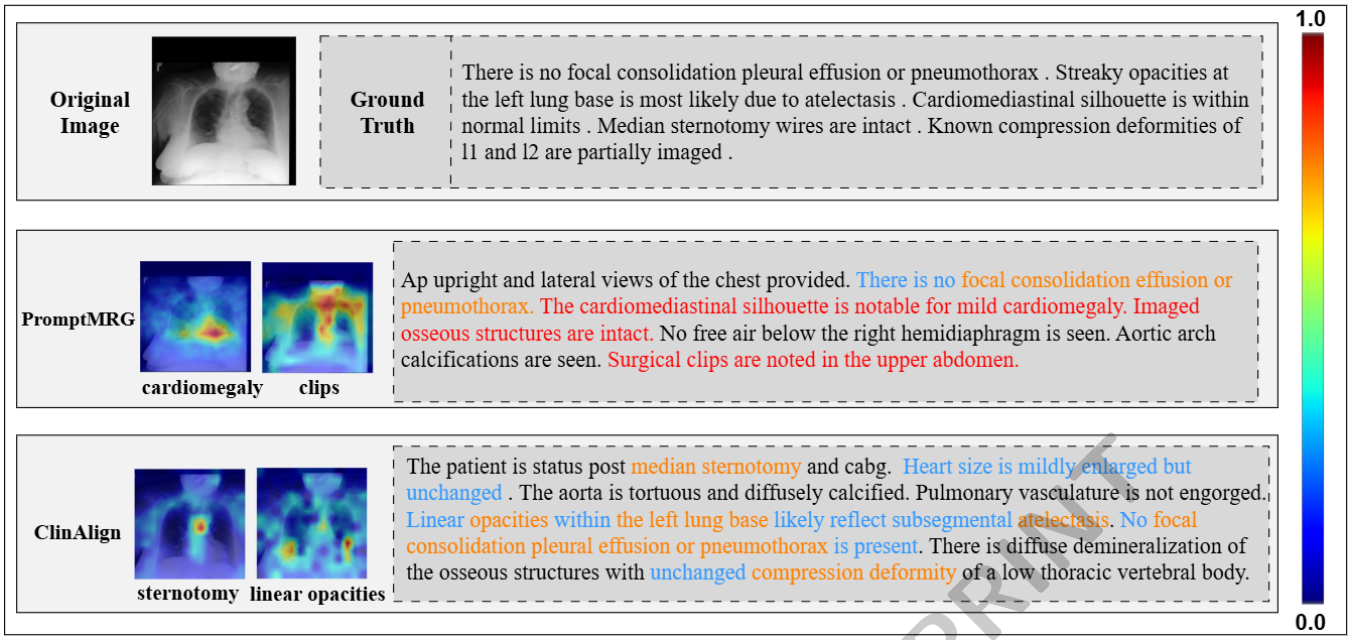


Figure 5: Qualitative comparison between PromptMRG and ClinAlign on a MIMIC-CXR case. Matched key entities are highlighted in orange, semantically consistent descriptions in blue, and unsupported statements absent from the ground truth in red.

retrieval enhancement and generation accuracy. Overall, appropriate λ and k are crucial for balancing semantic guidance, retrieval efficiency, and generation quality. Therefore, we adopt $\lambda = 4$ and $k = 500$ as the default settings in all subsequent experiments.

5 Visualization Case

To evaluate the effectiveness of ClinAlign in generating radiology reports, we randomly select one case from MIMIC-CXR for qualitative analysis. From figure 5, we can find that the report generated by ClinAlign performs better in terms of clinical consistency and detailed description, focusing its attention on diagnostically relevant areas. Specifically, ClinAlign correctly captures postoperative findings (e.g., *median sternotomy*) and subtle imaging patterns (e.g., *linear opacities*), while excluding other key abnormalities such as consolidation, pleural effusion, or pneumothorax. Notably, in this case, ClinAlign covers all key medical disease entities mentioned in the actual report, demonstrating improved entity-level coverage and consistency, consistent with the method's significant improvement in CE metrics. In contrast, PromptMRG fails to identify these critical postoperative and subtle findings, its attention is scattered to coarse or non-specific areas, and it generates unsupported statements (highlighted in red), such as emphasizing mild cardiac hypertrophy or surgical clips not mentioned in the actual report or lacking sufficient imaging evidence. These errors further reduce the reliability and clinical utility of the generated reports. Overall, the results indicate that ClinAlign achieves more accurate localization of abnormalities and more reliable clinical semantic expression, thereby improving the accuracy and specificity of radiological report generation.

6 Conclusion

We propose ClinAlign, a clinical workflow-aligned memory retrieval framework for RRG task. The framework achieves multimodal integration by jointly modeling visual pattern memory and disease-aware textual memory: the MPPA module enhances local feature representation at the visual level, while the CGPA module converts disease predictions into structured prompts to guide text retrieval and report generation. Extensive experiments demonstrate that ClinAlign achieves competitive performance across multiple evaluation metrics and effectively improves the clinical consistency and quality of generated reports. In future work, we plan to extend ClinAlign to temporal and longitudinal radiology data and integrate it with larger foundation models, so as to better capture disease progression and further improve robustness and generalization for real-world clinical deployment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62576064), Chongqing Natural Science Foundation for Innovative Development Joint Fund (Grant No.CSTB2025NSCQ-LZX0147) and Chongqing Key Laboratory of Precision Diagnosis and Treatment for Kidney Diseases.

References

[Alsentzer *et al.*, 2019] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78, 2019.

[Bu *et al.*, 2024] 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.

[Chen *et al.*, 2021] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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*, pages 5904–5914, 2021.

[Chen *et al.*, 2022] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5152–5161, 2022.

[Chen *et al.*, 2023] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal intervention for medical report generation. *arXiv preprint arXiv:2303.09117*, 2023.

[Demner-Fushman *et al.*, 2015] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

[Denkowski and Lavie, 2011] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Draelos and Carin, 2020] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

[Hou *et al.*, 2023] Xiaodi Hou, Zhi Liu, Xiaobo Li, Xingwang Li, Shengtian Sang, and Yijia Zhang. Mkcl: Medical knowledge with contrastive learning model for radiology report generation. *Journal of Biomedical Informatics*, 146:104496, 2023.

[Hou *et al.*, 2025] Xiaodi Hou, Xiaobo Li, Mingyu Lu, Simiao Wang, and Yijia Zhang. Rrg-mamba: Efficient radiology report generation with state space model. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7410–7418, 2025.

[Huang *et al.*, 2023] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818, 2023.

[Jin *et al.*, 2024] 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.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[Li *et al.*, 2023] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023.

[Li *et al.*, 2024a] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In *European Conference on Computer Vision*, pages 162–180. Springer, 2024.

[Li *et al.*, 2024b] Shiyu Li, Pengchong Qiao, Lin Wang, Munan Ning, Li Yuan, Yefeng Zheng, and Jie Chen. An organ-aware diagnosis framework for radiology report generation. *IEEE Transactions on Medical Imaging*, 43(12):4253–4265, 2024.

[Lin *et al.*, 2023] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Liu *et al.*, 2021] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762, 2021.

[Liu *et al.*, 2024a] Aohan Liu, Yuchen Guo, Jun-hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Medical Imaging*, 43(7):2657–2669, 2024.

[Liu *et al.*, 2024b] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18635–18643, 2024.

[Mao *et al.*, 2025] Po-Yuan Mao, Tien Hoang Nguyen, Wray Buntine, Mohammed Bennamoun, et al. Hope: A memory-based and composition-aware framework for zero-shot learning with hopfield network and soft mixture

of experts. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1101–1110. IEEE, 2025.

[Nooralahzadeh *et al.*, 2021] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, 2021.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Qin and Song, 2022] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022.

[Ramsauer *et al.*, 2020] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[Singh *et al.*, 2025] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025.

[Tanida *et al.*, 2023] 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.

[Tanno *et al.*, 2025] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.

[Tao *et al.*, 2024] Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4145–4156, 2024.

[Wang *et al.*, 2023a] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11558–11567, 2023.

[Wang *et al.*, 2023b] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

[Wang *et al.*, 2024a] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pages 183–200. Springer, 2024.

[Wang *et al.*, 2024b] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2199–2210, 2024.

[Wang *et al.*, 2024c] Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2gencsr: Retrieving context samples for large language model based x-ray medical report generation. *arXiv preprint arXiv:2408.09743*, 2024.

[Wang *et al.*, 2025] Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating associative disease-aware vision token memory for llm-based x-ray report generation. *arXiv preprint arXiv:2501.03458*, 2025.

[Xiao *et al.*, 2025] Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8664–8672, 2025.

[Xie *et al.*, 2023] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Medim: Boost medical image representation via radiology report-guided masking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–23. Springer, 2023.

[Yan and Pei, 2022] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2982–2990, 2022.

[Yang *et al.*, 2022] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.

[Yang *et al.*, 2023] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, Kevin S Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023.

[Yang *et al.*, 2024] Yan Yang, Jun Yu, Zhenqi Fu, Ke Zhang, Ting Yu, Xianyun Wang, Hanliang Jiang, Junhui Lv, Qingming Huang, and Weidong Han. Token-mixer: Bind image and text in one embedding space for medical image reporting. *IEEE Transactions on Medical Imaging*, 43(11):4017–4028, 2024.