Seeing Through Multiple Views: Parameter-Efficient Fine-Tuning via Selective Neurons for Consistent Radiology Report Generation

Yucheng Chen[1,2,3], Jinjing Zhu[4], Yang Yu[5], Yufei Shi[1,2], Hane Naghshbandi[1,2,3], Jinhua Liu[1,2], Angela S. Koh[6], Fang Fen[5], Kian Eng Ong[1], and Si Yong Yeo[1,2,3⋆]

[1] MedVisAI Lab
[2] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[3] Centre of AI in Medicine, Singapore
[4] The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong
[5] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[6] National Heart Centre Singapore (NHCS), Singapore

Abstract. Recent years have seen substantial advances in radiology report generation (RRG), yet existing approaches predominantly adopt direct feature fusion when handling multi-view X-ray images. Such approaches overlook the potential clinical inconsistencies and inaccuracies arising when a single model processes different views, adversely impacting performance and clinical reliability. To this end, we introduce **View-PNDF** (**View**-specific **P**attern **N**euron **D**etection and **F**ine-tuning), a parameter-efficient framework that fosters view-consistent report generation from a neuronal perspective. Specifically, View-PNDF comprises: (i) a view-specific neuron detection module identifying neurons responsive to particular views, (ii) a verification module quantifying the existence of these neurons, and (iii) a selective fine-tuning strategy strengthening detected neurons while preserving view-agnostic representations. By updating only view-specific neurons, View-PNDF achieves consistent diagnoses across different views with reduced computational costs. Subsequently, we employ Large Language Models (LLMs) to consolidate the view-specific reports into a complete radiology report. Furthermore, we use traditional Natural Language Generation (NLG) metrics-based assessment on integrated reports for baseline comparison and employ LLM-based assessment (e.g., GPT-4o) on view-specific reports to capture clinical significance. Extensive experiments on two medical RRG benchmarks demonstrate that View-PNDF substantially improves view-specific chest X-ray report generation quality while maintaining robust general-view performance.

Keywords: Radiology Report Generation · Vision-Language Model · Multi-view Consistency.

⋆ Corresponding author

1 Introduction

Radiology report generation (RRG) aims to automatically generate free-text descriptions from radiological images such as X-rays [19] and CT scans [11]. Current RRG approaches can be categorized into single-view and multi-view methods. Single-view models [7, 23, 28, 15, 21, 12] focus on generating reports from individual radiographic projections but suffer from inherent limitations in capturing comprehensive anatomical information. Multi-view approaches [6, 34, 22] address these limitations by incorporating information from multiple perspectives - such as frontal, lateral, posteroanterior (PA), and anteroposterior (AP) views - demonstrating empirical improvements over their single-view counterparts. This multi-view paradigm aligns with clinical practice, where radiologists routinely examine each imaging view individually and then integrate these per-view observations into a comprehensive report.

Multi-view medical imaging is clinically essential for identifying posterior abnormalities, localizing lesions, and detecting conditions obscured in single projections [13, 35, 36, 24]. However, existing multi-view RRG methods [6, 26, 34, 22] primarily employ fusion strategies to enhance feature representation, overlooking a critical issue: different views exhibit intrinsic variations that challenge automated systems. Such variations may lead to inconsistent interpretations across views, compromising diagnostic reliability and report quality. Current systems like InternVL-DA-Medical-4B [5] generate reports from frontal views that generally align with GT reports, while lateral view outputs often contain contradictory findings. Such inconsistencies pose significant clinical risks, particularly when frontal images are unavailable due to patient positioning constraints, trauma conditions, or equipment limitations, making lateral views the primary diagnostic input. In settings where frontal and lateral views are provided separately, inconsistent outputs may further lead to confusion, misinterpretation, and unreliable clinical decision-making. This raises a fundamental question: *what underlying mechanisms cause these view-specific discrepancies, and how can we systematically address them?* Without understanding the root causes at the model's internal level, solutions may only address symptoms rather than the core problem. Inspired by [3, 27], this motivates neuron-level analysis to identify which computational components are responsible for view-dependent variations, enabling targeted interventions to systematically improve cross-view consistency.

In this paper, we present a comprehensive neuron-level analysis of RRG models, namely **View**-specific **P**attern **N**euron **D**etection and **F**ine-tuning (**View-PNDF**). Rather than retraining the entire model, we first introduce a View-specific Neuron Detection (**VND**) module to identify a compact set of pattern neurons responsible for view-specific inconsistency. We then apply a View-specific Neuron Verification (**VNV**) module to quantitatively evaluate the contribution of these detected neurons by selectively deactivating them and measuring the impact on view-specific report generation. Building on these insights, we further propose a neuron-level adaptation strategy, View-specific Neuron Fine-tuning (**VNF**), which focuses training exclusively on the identified pattern neurons. The targeted fine-tuning enables the model to generate consistent view-specific findings from any available view, while findings from multiple views can be further consolidated into an integrated radiology report using LLMs. For evaluation, we use standard Natural Language Generation (NLG) metrics to assess the quality of the integrated reports. However, evaluating view-specific reports presents unique challenges as conventional NLG metrics fail to capture clinical validity: reports describing different anatomical views may receive similar scores even when they express contradictory clinical findings since NLG metrics measure textual overlap rather than medical semantics. To address this limitation, we employ a LLM–based evaluation method (e.g., GPT-4o) to assess the semantic fidelity and diagnostic accuracy of view-specific reports. By integrating view-aware methodology and appropriate evaluation strategies with RRG, our proposed framework offers a targeted and efficient solution for enhancing multi-view RRG performance.

2 Method

In this section, we introduce **View**-specific **P**attern **N**euron **D**etection and **F**ine-tuning (**View-PNDF**), a framework designed to enhance Radiology Report Generation (RRG) by identifying and optimizing neurons responsible for processing specific X-ray views (e.g., frontal vs. lateral). As illustrated in Fig. 1, our approach consists of three key components: View-specific Neuron Detection (**VND**), Neuron Verification (**VNV**), and Neuron Fine-tuning (**VNF**).

2.1 View-specific Pattern Neuron Detector (VND)

We hypothesize that within the RRG model - specifically in the linear and self-attention layers of the Transformer decoder - distinct subsets of neurons are specialized for handling different X-ray views. We define these as *View-specific Pattern Neurons*.

Neuron Influence via Perturbation. Ideally, the importance of a neuron can be measured by the impact of its deactivation on the model's output. Formally, let $\mathcal{RRG}(x, \theta)$ denote the generated embedding for a view-specific input sample x (comprising visual features and prompt tokens) given model parameters θ. The influence I of the k-th neuron at the l-th layer, denoted as n_l^k, on a specific view pattern v is defined as the perturbation effect:

$$I(n_l^k|x^v) = \left\|\mathcal{RRG}(x^v, \theta) - \mathcal{RRG}(x^v, \theta \setminus w_l^k)\right\|_2, \quad (1)$$

where $\theta \setminus w_l^k$ represents the model weights with the specific neuron n_l^k deactivated (zeroed out).

Efficient Detection via Attribution Scores. Since calculating Eq. 1 for every neuron across the entire dataset is computationally prohibitive, we follow prior work [8], which suggests that neurons with high activation values are strong indicators of pattern importance. We therefore employ an efficient *Attribution*

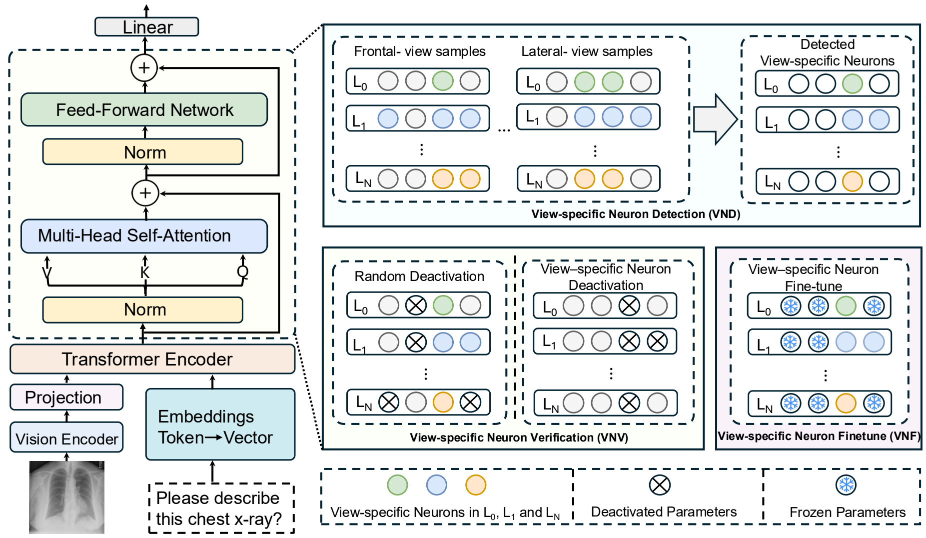

Fig. 1: The overall framework of View-PNDF, which consists of four components: the transformer structure in VLMs for RRG; View-specific Neuron Detection (VND) module identifies neurons associated with the specific views of chest X-rays; View-specific Neuron Verification (VNV) module verifies the detected pattern neurons; and View-specific Neuron Fine-tuning (VNF) module fine-tunes the detected view-specific neurons for RRG network.

Score as a proxy for detection. The attribution score for neuron n_l^k given input x^v is defined as the magnitude of its activation output:

$$\mathrm{Attr}(n_l^k|x^v) = \left\| \sum_t f(x^v, w_l^k)_t \right\|_1 , \tag{2}$$

where $f(\cdot)$ computes the activation values (e.g., the output of a specific linear projection or attention head) across the sequence dimension t.

Detection Strategy. To robustly identify neurons associated with a view v (e.g., lateral), we utilize a set of detection samples $\mathcal{D}_{detect} = \{x_1^v, \ldots, x_B^v\}$. A neuron is classified as a *View-specific Pattern Neuron* if it consistently ranks among the most active neurons across all samples in the detection set. The set of pattern neurons $\mathcal{N}^v$ is formally defined as:

$$\mathcal{N}^v = \bigcap_{b=1}^{B} \left\{ n_l^k \mid \mathrm{rank}(\mathrm{Attr}(n_l^k|x_b^v)) \leq \gamma N \right\}, \tag{3}$$

where $\mathrm{rank}(\cdot)$ sorts neurons in descending order of attribution score, N is the total number of neurons in the layer, and γ is a selection ratio (e.g., top 5%). This intersection strategy ensures we select neurons that are consistently salient for the specific view, filtering out noise from instance-specific variations.

2.2 Verification and Fine-tuning

View-specific Neuron Verification (VNV). To validate that the detected set $\mathcal{N}^v$ genuinely controls view-specific generation, we employ the perturbation metric from Eq. 1 on a held-out verification set. We compare the performance drop when deactivating the detected pattern neurons versus deactivating a random set of neurons of the same size. The existence of view-specific neurons is confirmed if:

$$\sum_{d=1}^{D} \mathcal{L}_{\Delta}(\mathcal{N}^{v}_{deact}) \gg \sum_{d=1}^{D} \mathcal{L}_{\Delta}(\mathcal{N}^{rand}_{deact}), \tag{4}$$

where $\mathcal{L}_{\Delta}$ represents the degradation in generation quality (e.g., increase in loss or decrease in clinical accuracy metrics) when neurons are deactivated.

View-specific Neuron Fine-tuning (VNF). Once identified and verified, we employ a targeted fine-tuning mechanism. Unlike standard fine-tuning methods which updates all parameters, VNF freezes the majority of the model and optimizes *only* the identified pattern neurons $\mathcal{N}^v$. This allows the model to enhance its representation of specific views without catastrophically forgetting general language generation capabilities. The optimization objective is the standard cross-entropy loss over the view-specific fine-tuning set $\mathcal{D}_{ft}$:

$$\mathcal{L}_{VNF} = -\sum_{(x,y)\in\mathcal{D}_{ft}} \sum_{t=1}^{T} \log P(y_t | y_{<t}, x; \theta_{\mathcal{N}^v}), \tag{5}$$

where $\theta_{\mathcal{N}^v}$ denotes that only the parameters of the pattern neurons are updated.

Table 1: NLG performance comparison of the integrated reports and reports generated by existing SOTA RRG methods on IU-XRay dataset. The subscripts MedGemma, InternVL, LLaVA, and Hulu denote that the VLM backbones used are MedGemma-4B, InternVL-DA-Medical-4B, LLaVA-Med-7B, and Hulu-Med-7B, respectively. BL: BLEU, MTR: METEOR, RG-L: ROUGE-L. Higher value indicates better performance.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
R2GEN [7]	0.390	0.293	0.236	0.161	0.187	0.371
R2GENCMN [6]	0.402	0.310	0.225	0.175	0.190	0.365
XPRONet [32]	0.480	0.312	0.225	0.175	0.190	0.364
RGRG [31]	0.373	0.249	0.175	0.126	0.168	0.264
M2KT [34]	0.489	0.315	0.228	0.174	–	0.379
CAMANET [33]	0.478	0.314	0.229	0.179	0.202	0.361
PromptMRG [15]	0.488	0.253	0.237	0.189	0.192	0.374
View-PNDF$_{\text{MedGemma}}$	0.496	0.322	0.240	0.194	0.205	0.385
View-PNDF$_{\text{InternVL}}$	0.492	0.318	0.234	0.191	0.198	0.382
View-PNDF$_{\text{LLaVA}}$	0.505	0.330	0.248	0.201	0.212	0.392
View-PNDF$_{\text{Hulu}}$	0.510	0.335	0.252	0.205	0.216	0.396
Average	**0.501**	**0.326**	**0.244**	**0.198**	**0.208**	**0.389**

Table 2: NLG and CE performance comparison of the integrated reports and reports generated by existing SOTA methods on MIMIC-CXR dataset. The subscripts MedGemma, InternVL, LLaVA, and Hulu denote that the VLM backbones used are MedGemma-4B, InternVL-DA-Medical-4B, LLaVA-Med-7B, and Hulu-Med-7B, respectively. BL: BLEU, MTR: METEOR, RG-L: ROUGE-L, CE: Clinical Efficacy, Prec.: Precision, Rec.: Recall.

METHOD	NLG Metrics						CE Metrics		
	BL-1	**BL-2**	**BL-3**	**BL-4**	**MTR**	**RG-L**	**Prec.**	**Rec.**	**F1**
R2GEN [7]	0.353	0.218	0.145	0.103	0.142	0.270	0.406	0.203	0.204
R2GENCMN [6]	0.348	0.206	0.135	0.094	0.136	0.266	0.440	0.325	0.374
XPRONet [32]	0.344	0.215	0.146	0.105	0.138	0.279	0.463	0.285	0.353
M2KT [34]	0.368	0.220	0.149	0.107	–	0.259	0.420	0.339	0.352
CAMANET [33]	0.367	0.219	0.151	0.107	0.142	0.279	0.467	0.323	0.381
PromptMRG [15]	0.369	0.223	0.142	0.099	0.147	0.260	0.471	0.198	0.352
MLRG [22]	0.398	0.236	0.159	0.156	0.169	0.278	0.491	0.398	0.401
View-PNDF$_{\mathrm{MedGemma}}$	0.408	0.245	0.166	0.164	0.174	0.286	0.502	0.415	0.408
View-PNDF$_{\mathrm{InternVL}}$	0.402	0.240	0.162	0.160	0.168	0.282	0.495	0.405	0.398
View-PNDF$_{\mathrm{LLaVA}}$	0.418	0.254	0.174	0.171	0.182	0.294	0.515	0.432	0.420
View-PNDF$_{\mathrm{Hulu}}$	0.425	0.260	0.179	0.176	0.188	0.301	0.522	0.445	0.428
Average	**0.413**	**0.250**	**0.170**	**0.168**	**0.178**	**0.291**	**0.509**	**0.424**	**0.414**

3 Experiments and Results

3.1 Dataset and Evaluation Metrics

Datasets. We evaluate on **IU-XRay** [9] and **MIMIC-CXR** [17], both containing frontal and lateral views. For IU-XRay (2,955 samples), we adopt the standard 70%/10%/20% split [16]. For MIMIC-CXR, to ensure balanced view representation, we curate a subset of 10,000 training samples and 1,000 validation/test samples, instead of the standard split [7], which is dominated by single-view reports.
Evaluation Metrics. Integrated reports are evaluated using NLG metrics, including BLEU [29], METEOR [10], and ROUGE-L [20]. For view-specific reports, given the limited ability of n-gram metrics to assess clinical validity, we employ GPT-4o [1] to score semantic fidelity on a scale of 1–10, following recent LLM-based evaluation practices [37].

3.2 Implementation

The model is trained using a PyTorch implementation on two NVIDIA A6000 GPUs. The initial learning rate is set as 4×10^{-6} with a cosine learning rate schedule. We use AdamW [25] optimizer to optimize the network with a weight decay of 0.5. We train the model for 1 epoch with a batch size of 8. The warm up ratio is set as 0.3.

Table 3: Performance comparison of medical VLMs with and without View-PNDF. Scores are evaluated by GPT-4o.

Model	**IU-XRay [9]**				**MIMIC-CXR [17]**			
	Frontal		**Lateral**		**Frontal**		**Lateral**	
	w/o	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**
MedGemma-4B [30]	7.1	7.3	2.9	6.8	6.2	6.3	3.6	5.9
InternVL-DA-Medical-4B [5]	8.5	8.6	3.7	7.9	7.4	7.6	2.9	7.1
LLaVA-Med-7B [18]	8.7	8.8	4.1	8.3	8.3	8.5	3.9	7.8
Hulu-Med-7B [14]	8.9	9.1	4.3	8.5	8.4	8.5	4.2	8.1

Table 4: NLG results of the InternVL-DA-Medical-4B on the IU-Xray dataset.

Before FT	**BL-1**	**BL-4**	**MTR**	**RG-L**
Frontal	0.398	0.156	0.163	0.265
Lateral	0.397	0.154	0.162	0.266
After FT	**BL-1**	**BL-4**	**MTR**	**RG-L**
Frontal	0.397	0.155	0.164	0.265
Lateral	0.397	0.154	0.163	0.264

Table 5: LLM evaluation scores of InternVL-DA-Medical-4B on the IU-Xray dataset.

Evaluator	**Frontal**	**Lateral**
GPT-4o	8.7	7.5
Qwen-72B	8.7	7.7
DeepSeek-V3	8.6	7.6

3.3 Comparative Experiment

We evaluated our method on IU-XRay [9] and MIMIC-CXR [17] against state-of-the-art methods. As shown in Tab. 1, applying our method to four VLM backbones (MedGemma, InternVL, LLaVA, Hulu) on IU-XRay consistently outperformed baselines across all metrics, with View-PNDF$_{\text{Hulu}}$ achieving the best results (e.g., 0.510 BLEU-1). On the MIMIC-CXR dataset (Tab. 2), our approach maintained this advantage, surpassing the leading baseline MLRG in both linguistic quality (ROUGE-L 0.291 vs 0.278) and Clinical Efficacy (F1 0.414 vs 0.401). These results confirm that fine-tuning view-specific neurons significantly enhances both the linguistic fluency and clinical accuracy of generated reports.

3.4 Ablation Study

To verify the effectiveness of each module, we conduct ablation studies on the IU-XRay [9] and MIMIC-CXR [17] datasets using the InternVL medical model [5].

The Effectiveness of View-PNDF. Tab. 3 details the GPT-4o evaluation of four VLMs with and without View-PNDF. The method consistently improved performance across all models and datasets. Notably, lateral view scores saw significant gains (e.g., +4.2 points on MIMIC-CXR for InternVL), demonstrating that View-PNDF effectively captures view-specific patterns to generate more accurate descriptions for both views.

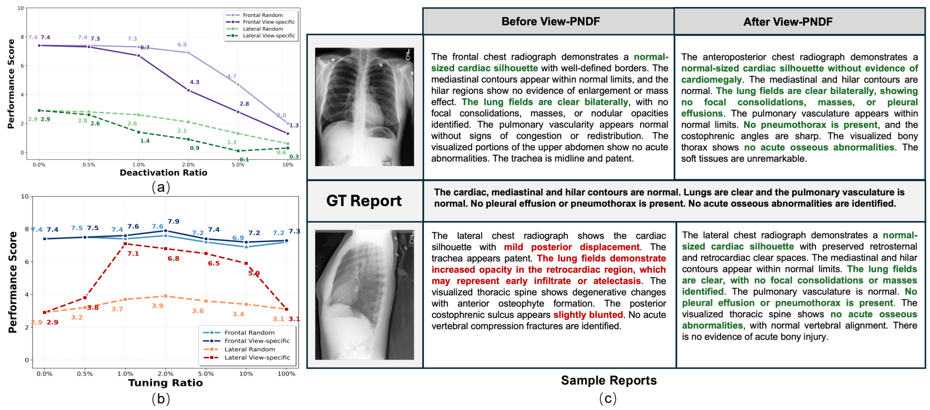

Fig. 2: Ablation studies of VND, VNV, and VNF on MIMIC-CXR dataset. (a) VND yields steeper performance drops than random deactivation, confirming neuron specificity. (b) VNF boosts lateral metrics while maintaining frontal quality. (c) Sample output from InternVL-DA-Medical-4B.

The Effectiveness of VND and VNV. To validate our detection mechanism, we compared view-specific versus random neuron deactivation (Fig. 2 (a)). View-specific deactivation caused drastic performance drops (e.g., lateral score dropping from 2.9 to 0.3 on MIMIC-CXR), whereas random deactivation had minimal impact. This significant disparity confirms that VND successfully identifies neurons critical for specific views.

The Effectiveness of VNF. Fig. 2 (b) shows that view-specific fine-tuning significantly boosts lateral view performance (e.g., reaching 7.3 vs. 3.7 for random tuning on MIMIC-CXR) while preserving or enhancing frontal view quality. This proves VNF effectively bridges the performance gap between views without compromising general capabilities.

LLM Evaluators. Tab. 4 shows that standard NLG metrics yield nearly identical scores before and after fine-tuning, failing to capture clinical nuances. In contrast, evaluations using GPT-4o [1], Qwen [2], and DeepSeek [4] (Tab. 5) reveal clear performance differences, demonstrating that LLM-based evaluation aligns better with clinical validity for this task.

Qualitative Results. Fig. 2 (c) illustrates the qualitative impact of View-PNDF. While the baseline model misses critical findings (e.g., sternal wires, cardiomegaly) and lacks inter-view consistency, the fine-tuned model successfully captures these pathologies across both frontal and lateral views. This confirms that our method enhances the detection of view-specific pathological patterns, yielding more comprehensive reports.

4 Conclusion

We proposed View-PNDF to detect and fine-tune view-specific pattern neurons in RRG models. Experiments confirm the existence of these neurons and show that optimizing a small fraction ($< 5.0\%$) significantly enhances generation consistency and accuracy across views. While demonstrated on chest X-rays, our parameter-efficient method holds potential for other modalities (e.g., CT, MRI). Future work will focus on extending this approach to multi-modal data and developing model-agnostic implementations.

Acknowledgments. This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 RS16/23. This Project is also partly supported by Centre of AI in Medicine, Singapore.

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
4. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
7. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
8. Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., Wei, F.: Knowledge neurons in pretrained transformers. arXiv preprint arXiv:2104.08696 (2021)
9. 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 (2015)
10. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation. pp. 85–91 (2011)

11. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
12. Huang, X., Chen, W., Liu, J., Lu, Q., Luo, X., Shen, L.: Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3769–3778 (2025)
13. Ittyachen, A.M., Vijayan, A., Isac, M.: The forgotten view: Chest x-ray-lateral view. Respiratory medicine case reports **22**, 257–259 (2017)
14. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
15. 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. vol. 38, pp. 2607–2615 (2024)
16. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274 (2020)
17. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
18. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
19. Liang, X., Zhang, Y., Wang, D., Zhong, H., Li, R., Wang, Q.: Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4967–4975 (2024)
20. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
21. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024)
22. Liu, K., Ma, Z., Kang, X., Li, Y., Xie, K., Jiao, Z., Miao, Q.: Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10348–10359 (2025)
23. Liu, Z., Zhu, Z., Zheng, S., Zhao, Y., He, K., Zhao, Y.: From observation to concept: A flexible multi-view paradigm for medical report generation. IEEE Transactions on Multimedia **26**, 5987–5995 (2023)
24. Loh, D.R., Yeo, S.Y., Tan, R.S., Gao, F., Koh, A.S.: Explainable machine learning predictions to support personalized cardiology strategies. European Heart Journal-Digital Health **3**(1), 49–55 (2022)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. Miao, Q., Liu, K., Ma, Z., Li, Y., Kang, X., Liu, R., Liu, T., Xie, K., Jiao, Z.: Evoke: Elevating chest x-ray report generation via multi-view contrastive learning and patient-specific knowledge. arXiv preprint arXiv:2411.10224 (2024)
27. Mu, J., Andreas, J.: Compositional explanations of neurons. Advances in Neural Information Processing Systems **33**, 17153–17163 (2020)

28. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
29. 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)
30. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
31. 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)
32. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. pp. 563–579. Springer (2022)
33. Wang, J., Bhalerao, A., Yin, T., See, S., He, Y.: Camanet: class activation map guided attention network for radiology report generation. IEEE Journal of Biomedical and Health Informatics **28**(4), 2199–2210 (2024)
34. 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)
35. Yang, X., Su, Y., Duan, R., Fan, H., Yeo, S.Y., Lim, C., Zhong, L., Tan, R.S.: Cardiac image segmentation by random walks with dynamic shape constraint. IET Computer Vision **10**(1), 79–86 (2016)
36. Yeo, S.Y., Xie, X., Sazonov, I., Nithiarasu, P.: Level set segmentation with robust image gradient energy and statistical shape prior. In: 2011 18th IEEE International Conference on Image Processing. pp. 3397–3400. IEEE (2011)
37. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)