RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment

Difei Gu[1], Yunhe Gao[1,2], Yang Zhou[1], Mu Zhou[1], and Dimitris Metaxas[1]

[1] Rutgers University
[2] Stanford University

Abstract. Medical image interpretation and report generation are essential for radiologists to identify and communicate observable findings of diseases. Major efforts in image-to-report generation require heavy language model training yet still suffer from producing reports with factual errors. In this study, we present RadAlign, demonstrating that a concept-based vision-language model can improve both predictive accuracy and report factual correctness without extensive language model training. Our key innovation is aligning visual features with medical diagnostic criteria in a shared representation space. Such alignment introduces core knowledge supervision and creates interpretable intermediate diagnosis results for LLMs to refine report generation. We also propose a cross-modal retrieval mechanism to provide additional clinical context of history cases for enhancing report generation accuracy. This unified approach achieves superior disease classification on MIMIC-CXR (average AUC: 0.885) and enables accurate report generation (GREEN score: 0.678 vs. SOTA: 0.634). RadAlign also demonstrates exceptional generalization capabilities, outperforming SOTA foundation and specialized models on the external OpenI dataset (AUC: 0.923 vs. 0.836). Code is available at https://github.com/difeigu/RadAlign.

Keywords: Vision-Language Model · Visual Concept Learning · Radiology Report Generation.

1 Introduction

Medical image interpretation and report generation play a vital role in the clinical workflow that can directly impact disease characterization and patient care [14]. For instance, chest radiograph interpretation [17] remains as a critical task, where clinicians must recognize subtle abnormalities and translate precise disease classifications into detailed reports. Accomplishing this complex task requires systematic efforts to capture a detailed state of the disease and generate comprehensive, well-reasoned explanations of these clinical findings [24].

Major research on chest radiographic interpretation falls into classification models [30,3,2] and image captioning approaches [4,11]. First, classification methods

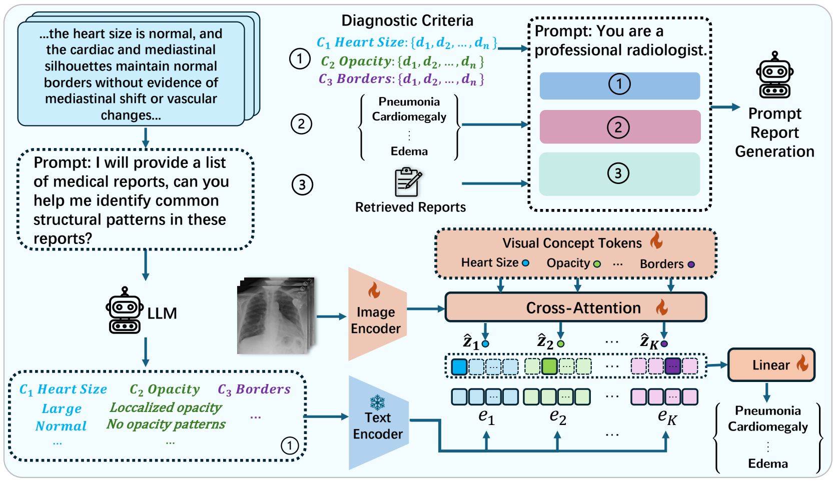
...the heart size is normal, and the cardiac and mediastinal silhouettes maintain normal borders without evidence of mediastinal shift or vascular changes...
Prompt: I will provide a list of medical reports, can you help me identify common structural patterns in these reports?
LLM
Diagnostic Criteria
1
C_1 Heart Size: $\{d_1, d_2, \dots, d_n\}$
C_2 Opacity: $\{d_1, d_2, \dots, d_n\}$
C_3 Borders: $\{d_1, d_2, \dots, d_n\}$
2
Pneumonia
Cardiomegaly
Edema
3
Retrieved Reports
Prompt: You are a professional radiologist.
1
2
3
Prompt Report Generation
Visual Concept Tokens
Heart Size
Opacity
...
Borders
Image Encoder
Cross-Attention
$\hat{z}_1$
$\hat{z}_2$
...
$\hat{z}_K$
Linear
Pneumonia
Cardiomegaly
Edema
C_1 Heart Size
C_2 Opacity
C_3 Borders
Large
Normal
Loccalized opacity
No opacity patterns
...
1
Text Encoder
e_1
e_2
...
e_K

Fig. 1: **Overview of the RadAlign framework.** Our unified VLM predicts three key components: **1. Diagnostic criteria** and associated concepts, formulated by an LLM based on historical reports, facilitate learning of image-concept alignment; **2. Disease prediction** is derived from this alignment, enabling an explainable classifier; **3. Augmented historical reports** are retrieved using learned visual concept tokens. These three components are combined and serve as knowledge-guided prompts for the LLM, ensuring a factually accurate report.

build on deep neural networks [9,2,30] and vision transformers [7,18] to show diagnostic precision in detecting pneumonia, cardiomegaly, and pulmonary edema. However, these models operate as black boxes by predicting only disease labels without explaining the visual semantic features that led to their predictions. This lack of interpretability limits their utility in real-world clinics. Second, growing efforts have investigated image captioning approaches [11,5,25,1] for generating free-text radiology reports. Despite of their advance, these methods often require extensive language model training yet still suffer from hallucinations - generating incorrect or unreliable information misaligned with the actual image content or medical knowledge [23,33].

The alignment between visual content and language context [8] is essential for developing human-level diagnostic reports. To illustrate, radiologists follow a structured process where they first assess specific diagnostic criteria and medical concepts (e.g., heart size, lung opacity, or pulmonary vessels) and then synthesize these observations with their medical knowledge to form detailed reports [10]. This clinical workflow motivates our development of **RadAlign**, a multi-modal framework that unifies the strength of predictive models with the reasoning capabilities of LLMs. Unlike prior approaches defining visual analysis and report generation as separate tasks [26], RadAlign purposely mirrors the radiologist's workflow on the concept-based image diagnosis. Our contributions are:

- A unified framework that bridges the gap between classification accuracy and detailed reporting through the vision-language concept alignment.
- A novel approach to medical report generation that mirrors radiologist workflow, combining visual feature recognition with LLM-based reasoning.
- A cross-modal retrieval-augmented generation system that enhances report reliability by grounding predictions in similar historical cases.
- Superior performance in classification (AUC: 0.885 on MIMIC-CXR, 0.923 on OpenI) and report generation benchmarks (GREEN score: 0.678 vs. SOTA: 0.634) with improved interpretability for clinical applications.

2 Related Work

Vision-Language Models (VLMs) seek to align visual and textual representations via contrastive learning and multimodal pre-training. General-purpose VLMs [22,3,13] trained on natural images often lack the specific medical knowledge required for disease understanding. Therefore, increased domain-specific adaptations have been explored including BioViL [3], MedCLIP [27] and MedKLIP [28]. These methods excel at learning joint representations to enable downstream tasks. They often focus on image-and-text matching in a pre-training framework, without explicitly considering patient information retrieval at inference time. In contrast, our effort emphasizes on learning domain-specific concepts, which serve as interpretable anchors for image analysis and textual reporting. This key procedure provides a transparent foundation for case retrieval and report generation, ensuring that the final report is clinically grounded.

Multimodel Caption Generation aims to integrate textual and visual information to improve the caption quality. Traditional approaches [4,16,21] employ neural networks to leverage both visual and textual features for generating coherent reports. However, these methods often function as black-box models, prioritizing the performance at the cost of interpretability. With the rise of large language models (LLMs), approaches like ChatCAD [26] aim to incorporate reasoning by combining classification networks, segmentation models, report generation modules, and LLMs. It is clear that these independent models could introduce significant computational overhead and increased integration complexity. Inconsistencies can emerge when aligning outputs from different modules, potentially compromising the final report's quality and reliability.

3 Methodology

3.1 Domain Knowledge Query

Inspired by how radiologists diagnose images, we first extract structured diagnostic criteria by mining expert-provided findings to create a foundation for concept-based diagnosis. Let $\mathcal{D} = \{(x, P, y)\}$ be our training set, where x is an image, P the ground-truth report findings, and $y \in \mathcal{Y}$ the disease label among N classes. We compile all findings $\mathcal{P} = \{P_1, P_2, \ldots, P_{|\mathcal{D}|}\}$ and prompt an LLM to derive a set of K disentangled diagnostic criteria $\{\mathcal{C}_i\}_{i=1}^{K}$. For example, for chest X-rays, the criteria might include `Heart Size`, `Lung Opacity`, `Diaphragm Position`, etc. We then query more detailed knowledge per criterion, grouping them by disease class as $\mathcal{C}_i = \{C_i^1, C_i^2, \ldots, C_i^{n_i}\}$. Each description indicates how that criterion manifests for each disease (e.g., `Heart Size` changes for *Cardiomegaly* but not for *Pneumonia*). Lastly, we build a mapping $f_m : \mathcal{C} \rightarrow \mathcal{Y}$ to link each concept description to one or more disease classes.This structured knowledge extraction provides crucial semantic anchors for our model to learn clinically relevant patterns.

3.2 Visual Concept Fine-grained Alignment

This component aims to discover and recognize specific visual features based on the structured diagnostic criteria, enabling the model to "see" like a radiologist. Using a pretrained vision-language model, we encode the textual criteria $\{\mathbf{e}_i\}_{i=1}^{K}$ via its text encoder $\mathcal{T}$. Each $\mathbf{e}_i \in \mathbb{R}^{n_i \times d}$ anchors the expert-derived concepts in embedding space. Meanwhile, we introduce K learnable visual concept tokens $\mathbf{z} \in \mathbb{R}^{K \times d}$ in the visual encoder $\mathcal{V}$. Given an image x, we extract features $\mathcal{V}(x)$ and use cross-attention to obtain:

$$\hat{\mathbf{z}} = \text{cross-attention}(\mathbf{z}, \mathcal{V}(x), \mathcal{V}(x)), \tag{1}$$

where $\mathbf{z}$ acts as the query. Each of the K tokens is encouraged to focus on the visual features pertinent to its corresponding criterion. To align visual and textual embeddings, we employ a domain-specific contrastive loss:

$$\mathcal{L}^{i}_{anchor}(\hat{\mathbf{z}}_i, \mathbf{e}_i) = -\log \frac{\exp\big(\text{sim}(\hat{\mathbf{z}}_i, \mathbf{e}_i^{\text{positive}})/\tau\big)}{\sum_{j=1}^{n_i} \exp\big(\text{sim}(\hat{\mathbf{z}}_i, \mathbf{e}_i^{j})/\tau\big)}, \tag{2}$$

where $\hat{\mathbf{z}}_i$ and $\mathbf{e}_i$ are matched concept embeddings, τ is a temperature parameter, and sim denotes dot-product similarity. This alignment process teaches the model to recognize clinically relevant patterns in radiographs, mimicking how radiologists diagnose with criteria.

3.3 Knowledge-guided Prompting

We propose a novel approach that leverages the diagnostic power of our aligned concept model without requiring extensive language model training. Our key insight is that the well-aligned concept-based vision-language model already contains sufficient diagnostic information for accurate report generation. Our vision-language model produces visual concept tokens $\hat{\mathbf{z}}_i$ aligned with diagnostic criteria anchors $\mathbf{e}_i$. We construct an explainable classifier using the similarity:

$$\hat{y} = W(\text{concat}(sim(\hat{\mathbf{z}}_1, \mathbf{e}_1), \ldots, sim(\hat{\mathbf{z}}_K, \mathbf{e}_K)))^{\intercal}, \tag{3}$$

where W represents the significance of each criterion's contribution toward classification. The total loss function combines cross-entropy for disease classification with the average contrastive loss:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{ce}}(\hat{y}, y) + \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}^{i}_{anchor}(\hat{\mathbf{z}}_i, \mathbf{e}_i). \tag{4}$$

To generate the report, we directly prompt the LLM with both the recognized criteria from our aligned model and the final classification prediction. Since the vision-language model is aligned in terms of diagnostic criteria concepts, it already contains the detailed findings necessary for accurate reporting. The LLM's role is primarily to reform these findings into a coherent, well-structured report rather than making diagnostic decisions. This approach uniquely combines the accuracy of our predictive model with the language capabilities of LLMs while significantly reducing hallucinations, as the factual diagnostic content is already ensured by the vision-language alignment.

3.4 Image-based Report Retrieval Augmentation (RAG)

While our aligned concept model provides accurate diagnostic findings, we recognize that general-purpose LLMs are not specifically trained for medical reporting. They require clinical context to understand appropriate radiology writing styles and terminology. Our image-based RAG system addresses this by providing relevant clinical examples that help the LLM reason more effectively for medical reporting. We construct a report database of training images:

$$\mathcal{Q} = \{(\hat{z}_i, P_i)\}_{i=1}^{|D|} \tag{5}$$

Where $(\hat{z}_i, P_i)$ is a key-value pair of visual concept tokens and corresponding reports. We precompute and store the visual concept tokens to minimize inference overhead. For each image $x \notin D$, we retrieve the most similar TopK cases:

$$\mathcal{P}_{\text{retrieve}} = \mathcal{Q}(\hat{z}_i, \hat{z}_i \in \text{TopK}_{\hat{z}\in\{\hat{z}_1,\hat{z}_2,\ldots,\hat{z}_{|D|}\}} sim(\hat{z}_i, \hat{\mathbf{z}}_{\mathbf{x}})) \tag{6}$$

Where $\hat{\mathbf{z}}_{\mathbf{x}}$ is the predicted concept token of any testing image. This retrieval mechanism grounds the LLM's output in validated clinical examples, helping maintain professional terminology and reporting conventions. By providing similar cases with confirmed diagnoses, we enable the LLM to better contextualize the aligned visual concepts into a clinically appropriate report. The classification results, concept findings, and retrieved reports are then incorporated into a unified prompt for the LLM, leveraging its editing capabilities while ensuring medical accuracy and relevance.

4 Experiment and Results

4.1 Experimental Setup

Dataset. We use **MIMIC-CXR** [12] for a comprehensive training and evaluation. The MIMIC-CXR dataset contains 377,100 chest X-ray images and corresponding radiology reports including findings, impressions, and patient history. We use five common classes including Atelectasis (AT), Cardiomegaly

Table 1: **Left**: Report generation comparison. **Right**: RadAlign with different LLMs.

Model	LLM	GREEN ↑
R2GenCMN	-	0.634
ChatCAD	4o-mini	0.633
ChatCAD	4o	0.634
RadAlign†	4o-mini	0.629
RadAlign††	4o-mini	0.648
RadAlign††	4o	**0.678**

† Initialized with ImageNet weights.
†† Initialized with BioViL weights.

RadAlign + LLM	GREEN ↑
ChatGPT 3.5-Turbo	0.648
ChatGPT 4o-mini	0.646
ChatGPT 4o	0.678
Claude 3.5-Sonnet	0.658
Llama 3.1	0.695

(CM), Consolidation (CD), Edema (ED) and Pleural Effusion (PE). To highlight RadAlign's out-of-domain generalization, we further evaluate on **IU X-ray (OpenI)**[6], which is unseen by all comparison methods. IU X-ray contains 7,470 chest X-ray images with corresponding reports from 3,955 patients.

Baselines. We evaluate our model against SOTA baselines for both disease classification and report generation tasks. For disease classification, we compare with: zero-shot foundation models, like CLIP [22], BiomedicalCLIP [31], and BioViL [3]. We also compare with domain-specific models that are trained on the MIMIC-CXR dataset, including PCAM [30]; ChatCAD [26], a multi-model integration LLM framework; LABO [29], an explainable VLM with concept bottleneck. For report generation, we compare against R2GenCMN [4], a cross-modal memory network for visual-textual integration, and ChatCAD [26].

Implementation Details. We prompt GPT-4 to query the diagnostic criteria. For our vision backbone, we conduct experiments using both ImageNet pretrained Resnet-50 weights and the BioViL CLIP Resnet-50 weights [3]. During training, we only optimize the visual encoder, visual concept tokens, and the final linear layer with AdamW optimizer for 40 epoches, using a decreasing learning rate of [1e-3, 1e-4], while keeping the text encoder fixed. All experiments are conducted using PyTorch with Nvidia RTX 8000 GPUs.

4.2 Results

Report Generation Comparison. We evaluate report generation quality using GREEN Score [19], a metric specifically designed for medical report assessment that leverages LLM-based reasoning to identify clinically significant errors. Unlike traditional metrics such as BLEU [20], ROUGE [15], and BERTScore [32] that only measure surface-level text similarity without considering factual correctness, GREEN focuses on accurately distinguishing between presence and absence of conditions and offers both quantitative scores and interpretable explanations that align well with expert judgment. For implementation details and more discussion, we refer readers to the original paper [19].

As shown in Table 1 (left), RadAlign achieves superior performance with a GREEN score of 0.678 using GPT-4o, substantially outperforming the baseline methods (0.634). Notably, we observe different scaling behaviors between methods: ChatCAD shows negligible improvement when upgrading from GPT-4o mini

Table 2: Classification results for different methods on F1 and AUC

(a) MIMIC-CXR

Model	AT		CM		CD		ED		PE		Average	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CLIP	0.200	0.507	0.200	0.540	0.000	0.497	0.060	0.498	0.200	0.500	0.132	0.508
BiomedCLIP	0.180	0.547	0.113	0.526	0.157	0.584	0.166	0.572	0.365	0.614	0.196	0.569
BioViL	0.388	0.705	0.431	0.715	0.165	0.806	0.329	0.783	0.582	0.769	0.379	0.756
PCAM*	0.618	0.838	0.628	**0.876**	0.432	0.787	0.514	0.868	0.755	0.937	0.589	0.861
ChatCAD	0.311	0.542	0.523	0.650	0.527	0.724	**0.641**	0.662	0.764	0.838	0.553	0.683
LABO	0.583	0.753	0.607	0.768	0.462	0.747	0.556	0.820	0.714	0.847	0.584	0.787
RadAlign†	0.628	0.841	0.650	0.873	**0.490**	**0.824**	0.616	0.916	0.779	**0.956**	**0.633**	0.882
RadAlign††	**0.634**	**0.853**	**0.653**	0.873	0.473	**0.824**	0.580	**0.924**	**0.820**	0.954	0.632	**0.885**

* MIMIC-CXR finetuned. † Initialized with ImageNet weights. †† Initialized with BioViL weights.

(b) IU X-ray (OpenI)

Model	AT		CM		CD		ED		PE		Average	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CLIP	0.272	0.517	0.249	0.559	0.000	0.499	0.046	0.502	0.101	0.500	0.134	0.515
BiomedCLIP	0.100	0.520	0.160	0.542	0.200	0.618	0.118	0.550	0.348	0.610	0.185	0.568
BioViL	0.272	0.517	0.249	0.559	0.000	0.499	0.046	0.502	0.101	0.500	0.134	0.515
PCAM*	0.540	0.569	0.715	0.862	**0.607**	**0.978**	0.505	0.809	**0.786**	0.961	0.630	0.836
RadAlign††	**0.695**	**0.851**	**0.737**	**0.913**	0.563	0.952	**0.618**	**0.934**	0.648	**0.963**	**0.652**	**0.923**

* MIMIC-CXR finetuned. †† Initialized with BioViL weights.

to GPT-4o, while RadAlign shows significant performance gains (0.648 to 0.678). This differential scaling highlights how RadAlign's unified vision-language alignment effectively leverages enhanced LLM reasoning capabilities based on recognized medical concepts, while ChatCAD's multi-model pipeline lacks alignment, introducing inconsistencies that limit the benefits of more powerful LLMs.

Classification Accuracy Comparison. Table 2 presents disease classification results in terms of F1 score and AUC. On MIMIC-CXR, RadAlign achieves the leading classification performance with an average F1 score of 0.633 and AUC of 0.885, outperforming both foundation models like BiomedCLIP, BioVil, and specialized methods like ChatCAD and PCAM. More impressively, when evaluated on the unseen OpenI dataset, RadAlign maintains strong performance with an average F1 score of 0.652 and AUC of 0.923, demonstrating excellent generalization capability and robustness to domain shifts.

Evaluation with different LLMs. We evaluated RadAlign with various large language models to assess its compatibility and generalizability,as shown in Table 1 (right). All tested LLMs, including ChatGPT (3.5, 4o-mini, 4o), Claude 3.5-Sonnet, and Llama 3.1 7B, achieved higher GREEN scores (0.646–0.695) than the previous baseline of 0.634, showing that our visual concept alignment approach is robust across different LLM architectures. More advanced models generally performed better (e.g., ChatGPT improved from 0.648 to 0.678).

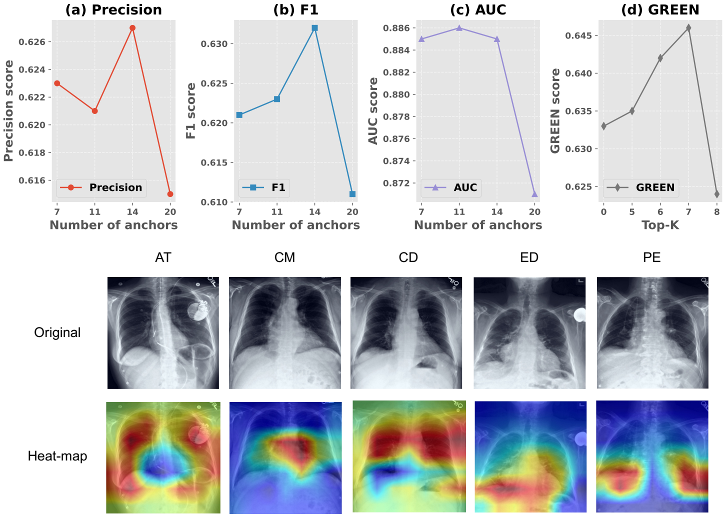

Fig. 2: **Top: Ablation studies**. (a-c) show classification performance of different number of concept anchors. (d) illustrates GREEN-score performance when varying the number of retrieved similar reports K. **Bottom: Visualization**. Attention map of concept tokens for different disease classes (AT, CM, CD, ED, PE), with warmer colors indicating higher attention scores.

Ablation Studies. Our experiments identified optimal parameters for RadAlign: performance peaked with 14 concept anchors across all metrics (Fig. 2, top a-c), as additional anchors introduced noise rather than meaningful features. For report retrieval, K=7 similar reports yielded the highest GREEN-Score (Fig. 2, top d), balancing sufficient semantic guidance without introducing misleading information from less relevant cases.

Concept Interpretation. RadAlign enables transparent interpretation of its decision-making process through visualization of concept token attention weights, displaying disease-specific localization patterns that align with clinical expertise. In Fig. 2 bottom, the attention heatmaps highlight anatomically-relevant regions for each condition. For example, for Atlectasis (AT), the heatmap highlights specific areas around the edge of the lung fields that are indicative of abnormalities, while for Cardiomegaly (CM), attention is drawn to distinct location at the heart region. These visualizations validate that our concept tokens can capture clinically meaningful features to assess the model's reasoning process.

5 Discussion and Conclusion

We introduce RadAlign, a novel framework that aligns visual features with medical concepts using a specialized Vision-Language Model. Unlike conventional methods that rely on extensive language model training or basic LLM prompting, RadAlign leverages a robust, concept-driven alignment strategy to map image features to diagnostic criteria. RadAlign achieves superior disease classification with an AUC of 0.885 on MIMIC-CXR and 0.923 on OpenI, while generating high-quality reports with a GREEN score of 0.678, outperforming state-of-the-art baselines. By integrating retrieval-augmented generation, RadAlign enhances factual accuracy and interpretability, drawing on similar historical cases to reduce hallucination. Notably, RadAlign bypasses extensive language model training, offering an efficient solution for clinical applications. Its concept-driven design ensures transparency by mirroring radiologists' diagnostic workflows.

Acknowledgments. This research has been partially funded by research grants to D. Metaxas through NSF: 2310966, 2235405, 2212301, 2003874, 1951890 and NIH 2R01HL127661.

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

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Asif, S., Wenhui, Y., Jin, H., Jinhai, S.: Classification of covid-19 from chest x-ray images using deep convolutional neural network. In: 2020 IEEE 6th international conference on computer and communications (ICCC). pp. 426–433. IEEE (2020)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
6. Demner-Fushman, D., Antani, S., Simpson, M., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gao, Y., Gu, D., Zhou, M., Metaxas, D.: Aligning human knowledge with visual concepts towards explainable medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hodler, J., Kubik-Huch, R.A., von Schulthess, G.K.: Diseases of the chest, breast, heart and vessels 2019-2022: diagnostic and interventional imaging (2019)

11. 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)
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
14. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018)
15. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
16. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
17. McBee, M.P., Awan, O.A., Colucci, A.T., Ghobadi, C.W., Kadom, N., Kansagra, A.P., Tridandapani, S., Auffermann, W.F.: Deep learning in radiology. Academic radiology **25**(11), 1472–1480 (2018)
18. Okolo, G.I., Katsigiannis, S., Ramzan, N.: Ievit: An enhanced vision transformer architecture for chest x-ray image classification. Computer Methods and Programs in Biomedicine **226**, 107141 (2022)
19. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
20. 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)
21. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 448–458 (2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
23. Ramesh, V., Chi, N.A., Rajpurkar, P.: Improving radiology report generation systems by removing hallucinated references to non-existent priors. In: Machine Learning for Health. pp. 456–473. PMLR (2022)
24. Reale-Nosei, G., Amador-Domínguez, E., Serrano, E.: From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. Medical Image Analysis p. 103264 (2024)
25. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
26. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)

28. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training in radiology. arXiv preprint arXiv:2301.02228 (2023)
29. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)
30. Ye, W., Yao, J., Xue, H., Li, Y.: Weakly supervised lesion localization with probabilistic-cam pooling. arXiv preprint arXiv:2005.14480 (2020)
31. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
32. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
33. Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D.: Data-centric foundation models in computational healthcare: A survey. arXiv preprint arXiv:2401.02458 (2024)