ORIGINAL RESEARCH Open access

Cross-Modal Retrieval for Radiology Report Generation from Chest X-Ray: A Framework Using Contrastive Learning Between Image Patches and Report Sentences

Omar Khalid[1*], Sara Nadeem[1], Bilal Farooq[2], Hina Saeed[1]

Abstract

Chest X-ray report generation is time-consuming and contributes to radiologist workload and burnout, motivating the need for AI systems that can reduce cognitive burden while preserving clinical accuracy. Although encoder-decoder models can generate reports from images, they often suffer from hallucinations, producing findings that are not present or missing real abnormalities due to lack of explicit grounding in evidence, making them unreliable for clinical use. To address this, we propose a cross-modal retrieval framework that generates reports by retrieving and assembling clinically validated sentences from existing radiology reports rather than generating text from scratch. The system uses contrastive learning to align chest X-ray image patches with report sentences in a shared embedding space, enabling retrieval of the most relevant clinical descriptions. A patch encoder extracts visual features, a sentence encoder represents report text, and a retrieval module identifies semantically matching sentences, which are then composed into a coherent final report. Because all outputs are sourced from real clinical reports, the method substantially reduces hallucinations while improving factual reliability and interpretability. This retrieval-based approach offers a scalable and safer alternative to generative models and can be evaluated on datasets such as MIMIC-CXR and CheXpert for clinical accuracy and retrieval performance.

Keywords Contrastive learning, Cross-modal retrieval, Radiology report generation, Chest X-ray, Vision-language alignment

*Correspondence:
Omar Khalid
omar.khalid@gmail.com
[1] Department of AI in Healthcare Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[2] Department of Clinical Intelligence Systems, Qatar University, Doha, Qatar

Introduction

The volume of chest X-ray examinations performed daily in hospitals worldwide places substantial demands on radiology departments, contributing to physician burnout and diagnostic errors [1, 2]. Radiologists must interpret each image and produce structured reports documenting findings, impressions, and recommendations, a process that becomes increasingly challenging as case volumes rise [3]. Artificial intelligence systems that can automatically generate draft reports have therefore attracted significant research interest as a means of reducing cognitive burden and improving workflow efficiency [4, 5].

Current state-of-the-art approaches to radiology report generation primarily employ encoder-decoder architectures, where a convolutional neural network encodes the chest X-ray and a recurrent or transformer-based decoder generates the report text from scratch [3, 4]. While these models can produce fluent and structurally coherent

reports, they suffer from a fundamental limitation: the tendency to hallucinate findings that are not present in the input image [6, 7]. This occurs because generative models learn statistical correlations between images and text but lack explicit mechanisms for grounding their outputs in verified evidence, leading to plausible but potentially dangerous statements [6, 8].

Retrieval-augmented generation offers a promising alternative paradigm that grounds output text in actual documents from a trusted database, thereby reducing hallucination by construction [6, 9]. Rather than generating sentences word-by-word, retrieval-based systems first identify relevant content from existing reports and then compose or paraphrase retrieved passages [10]. For medical applications where factual accuracy is paramount, the ability to trace each output sentence to a clinically validated source provides a critical safety advantage over pure generation [11].

This paper proposes a conceptual framework for radiology report generation based on cross-modal retrieval between chest X-ray image patches and report sentences, learned via contrastive pre-training [1, 12, 13]. The framework assumes access to a large database of paired chest X-rays and radiology reports, from which sentence-level embeddings are extracted and indexed [14]. For a new input image, the system retrieves the most relevant sentences from the database and composes them into a coherent report, eliminating the need for unsupervised text generation [15]. The remainder of this paper details the framework architecture, encoding components, contrastive learning objective, retrieval mechanism, and report composition strategy.

Background

Chest X-ray report structure

Chest X-ray radiology reports follow a standardized structure that has evolved to support efficient clinical communication and decision-making. A typical report contains several sections, including patient demographics, clinical indication, technique, comparison studies, findings, and impression, with the findings section comprising the majority of descriptive content [2, 16]. The findings section is typically organized anatomically, describing observations in the lungs, cardiac silhouette, mediastinum, hila, pleura, and bones in a systematic order that facilitates rapid review by referring physicians [3].

Individual sentences within the findings section tend to follow predictable patterns, describing the presence, location, and characteristics of specific abnormalities or the normal appearance of anatomical structures [4, 16]. For example, a sentence might state "There is bilateral interstitial opacities consistent with pulmonary edema" or "The cardiac silhouette is within normal limits." This sentence-level structure is particularly advantageous for retrieval-based generation, as it allows the system to treat individual clinical observations as retrievable units that can be combined to form complete reports [17].

Existing report generation methods

Early approaches to automated radiology report generation adopted the encoder-decoder paradigm widely used in image captioning, where a CNN extracts visual features from the chest X-ray and an RNN generates the report token by token [3, 8]. Subsequent work replaced RNNs with transformer decoders, achieving improved fluency and the ability to model longer-range dependencies in report text [4, 18]. Despite these architectural advances, encoder-decoder models consistently exhibit high hallucination rates, generating findings that have no correspondence to the input image while sometimes omitting salient abnormalities [6, 7].

The hallucination problem in medical report generation is particularly severe compared to natural image captioning because radiology reports contain dense, precise descriptions of multiple anatomical regions and abnormalities [3, 19]. A model that correctly describes lung fields might hallucinate cardiomegaly, or correctly identify a right-sided abnormality while fabricating a left-sided finding [11]. These errors arise because the decoder learns spurious correlations during training, such as associating certain image features with commonly co-occurring but absent findings [4, 20].

Cross-modal retrieval

Cross-modal retrieval refers to the task of retrieving relevant items from one modality (e.g., text) given a query from another modality (e.g., image), enabled by learning a shared embedding space where semantically similar items from different modalities are positioned close together [1, 21]. The CLIP framework demonstrated that contrastive learning on large-scale image-text pairs can produce powerful cross-modal representations, achieving strong zero-shot retrieval performance by maximizing agreement

between matched image-text pairs while pushing apart mismatched pairs [1, 22]. This approach has been widely adopted for vision-language tasks where alignment between modalities is essential [23, 24].

For medical applications, cross-modal retrieval offers a natural mechanism for grounding visual observations in textual descriptions, as each image patch depicting an anatomical region or abnormality can be mapped to sentences that describe similar findings [12, 21]. Unlike natural images where captions describe holistic scenes, chest X-rays contain multiple independent findings distributed across spatial locations, making patch-level retrieval particularly appropriate [15]. A retrieval-based report generation system can treat the task as a form of cross-modal search: given visual evidence from an image, find the textual descriptions that best match that evidence from a database of verified reports [9, 10].

Figure 1 shows the proposed framework, which replaces hallucination-prone free-text generation with a grounded retrieval pipeline by aligning chest X-ray patches and radiology report sentences within a shared contrastive embedding space before assembling a traceable final report.

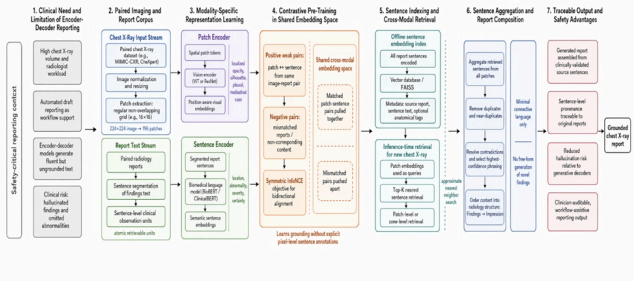

Figure 1. Cross-Modal Retrieval Architecture for Grounded Chest X-Ray Report Generation via Patch–Sentence Contrastive Alignment

Framework Overview

High-level architecture

The proposed framework operates in two phases: a pre-training phase that learns aligned cross-modal representations, and an inference phase that performs retrieval and composition [25]. During pre-training, a patch encoder processes chest X-rays by dividing them into non-overlapping spatial patches and producing embedding vectors for each patch, while a sentence encoder processes individual sentences from radiology reports to produce corresponding text embeddings [1, 13, 18]. A contrastive learning objective trains both encoders jointly, pulling patch embeddings close to the embeddings of sentences that describe those patches while pushing apart embeddings of unmatched patch-sentence pairs [14, 22].

During inference for a new chest X-ray, the patch encoder first computes embeddings for all patches in the image, and each patch embedding is used as a query to retrieve the most semantically similar sentence embeddings from a pre-computed index of the sentence database [12, 15]. The retrieved sentences are then passed to a composer module that aggregates them, removes duplicates, resolves contradictions, and arranges them into a coherent report following standard radiology structure [10]. Unlike generative models that produce text token by token with potential for error accumulation, this retrieval-based approach directly leverages clinically validated language from real reports [11, 17].

Core assumptions

The framework assumes access to a large, high-quality database of paired chest X-rays and radiology reports, such as MIMIC-CXR [6] or CheXpert [4], where each report is segmented into individual sentences that can be associated with specific spatial regions of the corresponding image [23]. This assumption enables weak supervision for patch-sentence alignment during pre-training, as the co-occurrence of a patch and a sentence within the same image-report pair provides a positive training signal even without explicit spatial annotations [5]. The database must be sufficiently large and diverse to ensure that for any new input image, relevant sentences can be retrieved with high precision [14].

Additionally, the framework assumes that radiology reports can be reliably segmented into sentences, each of which describes a coherent clinical observation or finding that can be meaningfully aligned with visual patterns in the image [2, 16]. This assumption holds for well-structured reports where sentences correspond to distinct anatomical regions or abnormality types, though reports with highly complex or multi-clause sentences may require more sophisticated segmentation [26]. The framework also assumes that the sentence database is static and clinically verified, meaning that retrieved sentences represent ground-truth observations from actual clinical practice [27].

Design principles

The framework is guided by three core design principles: grounding, factual accuracy, and controllable generation [9]. Grounding requires that every output sentence trace back to a source sentence from a real radiology report, providing a chain of evidence that can be inspected by clinicians [6, 11]. Factual accuracy demands that generated reports describe only findings that are actually present in the input image, which retrieval-based approaches satisfy by construction when retrieval precision is sufficiently high [26]. Controllable generation enables the system to produce reports of appropriate length and specificity, which can be achieved by adjusting the number of retrieved sentences per patch or applying confidence thresholds [7, 19].

These principles collectively address the primary failure mode of generative models: hallucination of non-existent findings [3, 8]. By shifting from generation to retrieval, the framework eliminates the need for the model to invent language or infer unobserved abnormalities, instead limiting output to statements that have been previously validated by radiologists [3, 4, 20]. The contrastive learning objective further enforces alignment between visual evidence and textual descriptions, ensuring that retrieved sentences genuinely correspond to what appears in the image rather than to spurious statistical correlations [1, 13, 22].

Table 1 identifies where factual risk enters the retrieval pipeline and shows how each design component functions as a specific control against hallucination, contradiction, redundancy, or omission.

Table 1. Failure-Mode–Control Alignment across the Cross-Modal Retrieval Pipeline for Radiology Report Generation

Pipeline stage	Primary technical role	Dominant failure mode	Clini conseq if uncont
Patch extraction	Decompose chest X-ray into localized visual units	Patch granularity too coarse or too fine; loss of anatomical context	Sub findings be miss anatom misloca
Patch encoder	Transform local image regions into discriminative visual embeddings	Embeddings capture texture but not clinically relevant semantics	Retrieva match v simila clinic inappro senter
Sentence segmentation	Define the atomic retrievable text units	Multi-clause sentences mix unrelated findings or qualifiers	Retrieva import part clinically statem distor mean
Sentence encoder	Represent report sentences in a clinically meaningful semantic space	Semantic compression may blur distinctions such as certainty, severity, or laterality	Wrong f may retrie desp superfi simi word
Contrastive pre-training	Align visual and textual representations	Weak supervision may align a patch with report sentences not truly tied to that region	Cross-r groun becor nois lower downst retrie preci
Retrieval index	Support scalable nearest-neighbor sentence search	Approximate retrieval may prioritize embedding proximity over true clinical appropriateness	False-p senter may ent candi poo
Patch-to-sentence retrieval	Select candidate report sentences for each visual unit	High-confidence retrieval of common normal statements overwhelms	Omissi subtl uncom abnorm

		rare abnormal findings	
Sentence aggregation	Merge retrieved sentences across patches	Duplicate retrievals and repetitive normal statements dominate the report	Rep becor verbo redun and l clinic usab
Contradiction resolution	Reconcile inconsistent retrieved statements	Coexistence of mutually exclusive findings (e.g., normal heart size vs cardiomegaly)	Major fa inconsis and lo clinician
Composer	Order validated sentences into final report structure	Over-composition or excessive editing reintroduces generative drift	Groun evidenc be refra into l faith langu
Final output	Deliver clinically usable draft report	Retrieved report is traceable but incomplete	Omiss errors pers desp reduc hallucir

Image Patch Encoding

Patch extraction

The patch encoder first divides the input chest X-ray into a regular grid of non-overlapping spatial patches, each of size 16×16 pixels, which preserves the spatial organization of anatomical structures while providing sufficiently granular visual information for localized finding detection [15, 18]. For a standard chest X-ray of size 224×224, this yields 196 patches, each corresponding to a specific anatomical region such as the right upper lung zone, left hilum, or cardiac silhouette [25]. The patch-based representation enables the framework to perform retrieval at the level of individual clinical observations rather than treating the entire image as a monolithic entity [17, 21].

Patch encoder architecture

Each extracted patch is independently processed by a vision encoder, which can be implemented either as a Vision Transformer (ViT) that treats each patch as a token with added positional embeddings, or as a residual convolutional network (ResNet) that processes patches in parallel [18, 25]. The encoder outputs a fixed-dimensional embedding vector for each patch, capturing visual features such as opacity patterns, silhouette abnormalities, and calcification that are relevant for radiological interpretation [22, 23]. Positional information is preserved either through the positional encoding in ViT or through the spatial arrangement of patch indices, ensuring that the framework can distinguish between similar findings in different anatomical locations [15, 24].

Sentence Encoding

Sentence segmentation

Each radiology report in the database is segmented into individual sentences using rule-based or learned sentence boundary detection, with special attention to clinical abbreviations and numbered lists that may appear in findings sections [2, 16]. Sentence segmentation typically yields 10-30 sentences per report, each representing a distinct clinical observation such as "There is left lower lobe consolidation" or "No pneumothorax is identified" [26]. The segmentation granularity is critical because sentences serve as the atomic retrievable units in the framework; overly long sentences may contain multiple unrelated findings, while overly short sentences may fragment clinically meaningful observations [17, 27].

Sentence encoder architecture

Each segmented sentence is encoded using a pretrained biomedical language model, such as BioBERT or ClinicalBERT, which have been fine-tuned on radiology text corpora to capture domain-specific semantics and terminology [16, 28, 29]. More recent models like RadBERT [16] and MedKLIP [24] have further advanced domain-specific language representation. The sentence encoder outputs a fixed-dimensional embedding vector that

represents the semantic content of the sentence, including the anatomical location, abnormality type, severity modifiers, and certainty expressions [9, 26]. Unlike generic sentence encoders, biomedical variants are sensitive to clinical distinctions such as the difference between "infiltrate" and "consolidation" or "possible" versus "definitive," which are crucial for accurate radiology reporting [24, 28].

Contrastive Pre-Training

Contrastive objective

The contrastive pre-training phase learns to align patch embeddings with sentence embeddings using the InfoNCE loss, which maximizes the cosine similarity between positive patch-sentence pairs while minimizing similarity between negative pairs drawn from different combinations [1, 13, 22]. A positive pair consists of an image patch and a sentence from the same report, under the weak assumption that the sentence describes some finding visible somewhere in the image, even if the exact spatial correspondence is unknown [14, 18]. Negative pairs are formed by matching patches with sentences from different reports, or from non-corresponding regions within the same report, providing a rich set of comparisons that push apart semantically mismatched representations [5, 15].

The training objective is computed symmetrically, with both patch-to-sentence and sentence-to-patch contrastive losses, encouraging the embedding space to be well-structured for bidirectional retrieval [1, 22]. For a batch of B image-report pairs, the loss encourages each patch embedding to be close to all sentence embeddings from its paired report while being distant from sentence embeddings from other reports [13, 21]. This weak supervision strategy is essential because radiology reports do not typically contain spatial annotations linking specific sentences to specific image regions, yet the overall co-occurrence signal provides sufficient alignment for retrieval [25, 27].

Training data

Training data for contrastive pre-training consists of paired chest X-rays and radiology reports from large publicly available datasets, primarily MIMIC-CXR [5] and CheXpert [4], which together provide hundreds of thousands of image-report pairs with diverse findings and patient populations [20]. Additional datasets and pre-training strategies such as GLoRIA [21], MedAug [5], and CXR-CLIP [22] have demonstrated the effectiveness of contrastive learning for chest X-ray interpretation. Each report is pre-processed by sentence segmentation, and each image is pre-processed by patch extraction, creating a dataset of patch-sentence pairs where the positive relationship is defined by co-occurrence within the same original image-report pair [14, 23]. The framework does not require explicit bounding boxes or pixel-level annotations, making it scalable to large datasets where such annotations are unavailable [15, 24].

The weak supervision assumption that all sentences in a report describe some finding in the paired image is approximately true for radiology reports, as radiologists describe only what they observe in the image and do not typically include extraneous information [2, 26]. However, this assumption may be violated for reports that describe prior studies, technical factors, or recommendations, which are typically excluded from sentence segmentation [27]. By filtering out non-findings sentences during pre-processing, the framework improves the signal-to-noise ratio of the positive pairs, leading to more accurate cross-modal alignment [9, 19].

Cross-Modal Retrieval

Building retrieval index

After pre-training, all sentences from the database are encoded using the trained sentence encoder, and their embeddings are stored in a vector index optimized for fast approximate nearest neighbor search, such as FAISS [12, 15]. The index also stores metadata for each sentence, including the source report identifier, the sentence text, and optionally the anatomical region or finding type if such annotations are available [10, 17]. Building the index is a one-time offline operation that can be performed on a large-scale computing cluster, after which the index can be loaded into memory for efficient online retrieval [9, 11].

Retrieval for new image

For a new chest X-ray presented at inference time, the patch encoder computes embeddings for all image patches, and each patch embedding is used as a query against the sentence index to retrieve the top-K most semantically similar sentences [1, 12, 22]. Retrieval can be performed independently for each patch, allowing different anatomical regions to retrieve different sentences that

describe localized findings [21, 25]. Alternatively, the framework can aggregate patch embeddings within anatomically defined zones (e.g., left lung, right lung, cardiac region) and perform zone-level retrieval to reduce redundancy [15, 24]. The retrieved sentences can be further refined using knowledge graph embeddings as proposed by Zhang *et al.* [12] and Yang *et al*. [10].

Report Composition

Sentence aggregation

After retrieval, the framework aggregates the set of retrieved sentences from all patches, removing exact duplicates and near-duplicates using semantic similarity thresholds to avoid repetitive statements such as multiple retrievals of "The lungs are clear" [6, 7]. Sentences are then grouped by anatomical region or finding type based on either explicit metadata tags or implicit semantic clustering, enabling the composer to organize the report following standard radiology structure [17, 27]. For regions where multiple retrieved sentences describe the same finding with different phrasings, the framework selects the sentence with highest retrieval confidence or the most specific clinical description [10, 11].

The aggregation process must also handle potential contradictions among retrieved sentences, such as one sentence stating "Cardiomegaly is present" while another states "The cardiac silhouette is normal" [4, 20]. Contradictions can arise when different patches retrieve inconsistent sentences due to ambiguous visual features or when the database contains conflicting reports for similar images [8, 19]. The framework can resolve contradictions by preferring sentences retrieved with higher confidence, by considering the frequency of findings in the database, or by flagging contradictions for clinician review rather than attempting automatic resolution [3, 26].

Generation with retrieved sentences

The final report is generated by feeding the aggregated and de-duplicated set of retrieved sentences into a decoder that arranges them into a coherent narrative following standard report structure: indication, comparison, findings, and impression sections [3, 6, 18]. Unlike generative decoders that produce text token-by-token, this decoder is primarily a composer that orders and formats existing sentences, potentially adding minimal connective phrases such as "Additionally," or "In comparison to prior examination," to improve fluency [9, 11]. The impression section can be generated by abstracting the key findings from the retrieved sentences, for example by selecting sentences that describe clinically significant abnormalities rather than normal findings [17, 27].

Because the composer does not generate novel clinical statements, it cannot introduce hallucinations beyond those already present in the retrieved sentences, which are themselves from verified reports [4, 7]. However, the framework must ensure that retrieved sentences are appropriate for the input image, which depends entirely on the quality of cross-modal retrieval [13, 14]. If retrieval precision is low, the composer may assemble a report that describes findings not present in the image, but such errors are traceable to specific retrieved sentences and can be audited by clinicians, unlike the opaque errors of generative models [8, 20].

Evaluation Strategy

Automatic metrics

Evaluation of the framework should prioritize clinical accuracy over surface-level text similarity, using metrics such as CheXbert [28] and RadGraph that assess the factual correspondence between generated reports and ground-truth annotations [4, 6, 20]. CheXbert [8] extracts structured labels for 14 common chest X-ray findings from generated reports, enabling comparison against reference labels to compute precision, recall, and F1 for each finding type. Hallucination rate, defined as the proportion of generated findings that are absent from the ground truth, is a critical safety metric that retrieval-based approaches are expected to substantially reduce compared to generative baselines [3, 7, 19].

Retrieval precision can be evaluated independently by measuring whether retrieved sentences for a given patch correspond to findings actually present in that patch's spatial location, using datasets with pixel-level abnormality annotations if available [12, 15]. The framework should also report coverage, defined as the proportion of ground-truth findings that are represented by at least one retrieved sentence, to ensure that retrieval does not omit clinically important abnormalities [10, 11]. Traditional natural language generation metrics such as BLEU and ROUGE are less informative for clinical applications because they reward fluency and n-gram overlap rather than factual correctness [3, 20].

Clinical validation

Beyond automatic metrics, the framework requires validation by practicing radiologists who can assess the clinical acceptability, factual accuracy, and usefulness of generated reports compared to encoder-decoder baselines [4, 6, 7]. A sample of test cases should be presented to multiple radiologists blinded to whether reports were generated by retrieval or by generative models, with each report rated on dimensions including "no clinically significant errors," "minor errors not affecting management," and "major errors that could harm patients" [8, 11]. Radiologists should also evaluate whether the retrieval-based reports are more or less useful as draft reports that could be edited rather than written from scratch [16, 27].

The clinical validation study should specifically measure the reduction in hallucination rates achievable with retrieval-based generation, comparing the proportion of reports containing at least one fabricated finding across frameworks [3, 4, 20]. If the retrieval framework achieves substantially lower hallucination rates while maintaining acceptable coverage of true findings, it would provide strong evidence for the clinical viability of retrieval-augmented approaches [9, 24]. However, validation must also assess the risk of omission errors, where the retrieval framework fails to retrieve sentences for subtle or rare findings that a generative model might occasionally describe correctly [7, 19].

Conclusion

This paper has presented a conceptual framework for radiology report generation from chest X-rays based on cross-modal retrieval between image patches and report sentences, learned through contrastive pre-training. The framework shifts away from the dominant encoder-decoder paradigm that generates reports from scratch, instead retrieving and composing clinically validated sentences from a database of real reports. By grounding each output sentence in a verifiable source, the framework directly addresses the critical safety limitation of generative models: their tendency to hallucinate non-existent findings.

The key advantages of the retrieval-based approach include inherent factual grounding, reduced hallucination rates, traceable provenance for each output sentence, and the ability to leverage large collections of existing reports without requiring explicit spatial annotations. The contrastive learning objective ensures that visual patterns in image patches are aligned with the linguistic descriptions that radiologists actually use, enabling precise cross-modal retrieval. The composer module then assembles retrieved sentences into coherent reports following standard radiology structure, producing outputs that inherit the clinical validity of source documents.

Limitations of the framework include its dependence on a comprehensive and representative sentence database, as retrieval precision will suffer for findings that are rare or absent from the training corpus. Retrieval latency may also be higher than pure generation, though approximate nearest neighbor indexes can achieve millisecond-scale search for databases of hundreds of thousands of sentences. Furthermore, the weak supervision assumption that all sentences in a report correspond to findings in the paired image may introduce noise into contrastive pre-training, potentially requiring filtering or weighting strategies to exclude non-findings sentences.

Future work should implement this framework using publicly available datasets MIMIC-CXR and CheXpert, evaluating retrieval precision, hallucination rates, and clinical acceptability against strong encoder-decoder baselines. The framework could be extended to incorporate temporal information from prior examinations, retrieving sentences that describe changes from previous studies. If validated, this retrieval-based approach offers a pathway toward clinically deployable report generation systems that reduce radiologist workload while maintaining the factual accuracy essential for patient safety.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 21 Oct 2023 Revised: 22 Dec 2023 Accepted: 06 Mar 2024
Published online: 20 July 2024

References

Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. Proc Annu Meet Assoc Comput Linguist. 2018;56:2577-86.

Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. Proc IEEE Conf Comput Vis Pattern Recognit. 2018:9049-58. https://doi.org/10.1109/CVPR.2018.00943.

Li CY, Liang X, Hu Z, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. Proc AAAI Conf Artif Intell. 2019;33(01):6666-73. https://doi.org/10.1609/aaai.v33i01.33016666.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell. 2019;33(01):590-7. https://doi.org/10.1609/aaai.v33i01.3301590.

Vu YN, Wang R, Balachandar N, Liu C, Ng AY, Rajpurkar P. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. Mach Learn Healthc Conf. 2021;755-69. https://doi.org/10.48550/arXiv.2102.10663.

Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0324-0.

Najdenkoska I, Zhen X, Worring M, Shao L. Uncertainty-aware report generation for chest X-rays by variational topic inference. Med Image Anal. 2022;82:102603. https://doi.org/10.1016/j.media.2022.102603.

Wang Z, Tang M, Wang L, Li X, Zhou L. A medical semantic-assisted transformer for radiographic report generation. Lect Notes Comput Sci. 2022;13439:655-64. https://doi.org/10.1007/978-3-031-16443-9_63.

Boecking B, Usuyama N, Bannur S, Castro DC, Schwaighofer A, Hyland S, et al. Making the most of text semantics to improve biomedical vision–language processing. Lect Notes Comput Sci. 2022;13697:1-21. https://doi.org/10.1007/978-3-031-20059-5_1.

Yang S, Wu X, Ge S, Zheng Z, Zhou SK, Xiao L. Radiology report generation with a learned knowledge base and multi-modal alignment. Med Image Anal. 2023;86:102798. https://doi.org/10.1016/j.media.2023.102798.

Tanida T, Müller P, Kaissis G, Rueckert D. Interactive and explainable region-guided radiology report generation. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2023;7433-42. https://doi.org/10.1109/CVPR52729.2023.00718.

Zhang Y, Wang X, Xu Z, Yu Q, Yuille A, Xu D. When radiology report generation meets knowledge graph. Proc AAAI Conf Artif Intell. 2020;34(07):12910-17. https://doi.org/10.1609/aaai.v34i07.6991.

Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. Mach Learn Healthc Conf. 2022;2-25. https://doi.org/10.48550/arXiv.2010.00747.

Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive learning from unpaired medical images and text. Proc Empir Methods Nat Lang Process. 2022;3876-87.

Wang F, Zhou Y, Wang S, Vardhanabhuti V, Yu L. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Adv Neural Inf Process Syst. 2022;35:33536-49. https://doi.org/10.48550/arXiv.2205.14044.

Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4(4):e210258. https://doi.org/10.1148/ryai.210258.

Yan S, Cheung WK, Chiu K, Tong TM, Cheung KC, See S. Attributed abnormality graph embedding for clinically accurate X-ray report generation. IEEE Trans Med Imaging. 2023;42(8):2211-22. https://doi.org/10.1109/TMI.2023.3247290.

Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. Proc Empir Methods Nat Lang Process. 2020;1439-49.

Nicolson A, Dowling J, Anderson D, Koopman B. Longitudinal data and a semantic similarity reward for chest X-ray report generation. Inform Med Unlocked. 2024;50:101585. https://doi.org/10.1016/j.imu.2024.101585.

Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns. 2023;4(9):100802. https://doi.org/10.1016/j.patter.2023.100802.

Huang SC, Shen L, Lungren MP, Yeung S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. Proc IEEE/CVF Int Conf Comput Vis. 2021;3942-51. https://doi.org/10.1109/ICCV48922.2021.00392.

You K, Gu J, Ham J, Park B, Kim J, Hong EK, et al. CXR-CLIP: Toward large scale chest X-ray language-image pre-training. Lect Notes Comput Sci. 2023;101-11. https://doi.org/10.1007/978-3-031-43999-5_10.

Zhou HY, Chen X, Zhang Y, Luo R, Wang L, Yu Y. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nat Mach Intell. 2022;4(1):32-40. https://doi.org/10.1038/s42256-021-00425-5.

Wu C, Zhang X, Zhang Y, Wang Y, Xie W. MedKLIP: Medical knowledge enhanced language-image pre-training for X-ray diagnosis. Proc IEEE/CVF Int Conf Comput Vis. 2023;21372-83. https://doi.org/10.1109/ICCV51070.2023.01958.

Zhou HY, Lian C, Wang L, Yu Y. Advancing radiograph representation learning with masked record modeling. arXiv preprint arXiv:2301.13155. 2023

Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun. 2023;14(1):4542. https://doi.org/10.1038/s41467-023-40260-7.

Bannur S, Hyland S, Liu Q, Perez-Garcia F, Ilse M, Castro DC, et al. Learning to exploit temporal structure for biomedical vision-language processing. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2023;15016-27. https://doi.org/10.1109/CVPR52729.2023.01444.

Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren M. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. Proc Empir Methods Nat Lang Process. 2020;1500-19.

Chen Z, Shen Y, Song Y, Wan X. Cross-modal memory networks for radiology report generation. Proc Annu Meet Assoc Comput Linguist Int Jt Conf Nat Lang Process. 2021;5904-14.