ORIGINAL RESEARCH **Open access**

Multimodal Vision-Language Model for Joint Interpretation of Chest X-Ray Images and Free-Text Radiology Requests to Generate Structured Preliminary Reports

Hassan Rahman[1*], Tariq Mahmood[1], Ali Raza[2]

Abstract

Chest X-ray is a commonly used imaging tool in both acute and routine care, but the increasing reporting workload highlights the need for structured preliminary reports that aid triage, reduce delays, and ensure clinical relevance. Current AI systems often focus on classification or generic report generation, neglecting critical factors like free-text radiology requests, clinical history, and comparison context, leading to reports that, while technically fluent, are insufficiently focused. This article proposes a multimodal vision-language model that interprets both chest X-ray images and free-text radiology requests to generate structured preliminary reports directly addressing the clinical question. The model combines a radiographic encoder based on vision transformers, a text encoder for requests and prior reports, a cross-modal attention module, and a structured report decoder, organizing the output into relevant sections such as indication, technique, findings, impression, comparison, and answer-to-request. By aligning report generation with the clinical request, the model ensures that it answers specific questions—such as concerns about pneumonia, pulmonary oedema, or pneumothorax—improving report relevance, reducing misinterpretation, and supporting safer human-in-the-loop review. However, its effectiveness relies on accurate alignment, factual consistency, uncertainty management, and validation in real-world radiology settings.

Keywords Radiology report generation, Chest X-ray, Vision-language model, Structured reporting, Free-text radiology request, Preliminary report

*Correspondence:
Hassan Rahman
hassan.rahman@gmail.com
[1] Department of Artificial Intelligence in Healthcare, University of Karachi, Karachi, Pakistan
[2] Department of Intelligent Clinical Analytics, National University of Sciences and Technology, Islamabad, Pakistan

Introduction

Chest X-ray imaging remains the most frequently performed radiological examination across emergency departments, inpatient wards, intensive care units, and outpatient clinics worldwide. Its low cost, portability, and rapid acquisition make it indispensable for triage and initial management of cardiopulmonary conditions, from pneumonia and heart failure to pneumothorax and post-operative complications. However, the volume of chest X-ray studies routinely exceeds the interpretive capacity of radiology services, particularly overnight, at weekends, or in under-resourced settings. Even in well-staffed academic centres, turnaround times for final radiologist reports can extend from hours to days, creating a bottleneck that directly impacts clinical decision-making. A patient with suspected tension pneumothorax may wait for confirmation before chest tube placement; a post-operative patient with new hypoxemia may remain without a definitive diagnosis while the imaging study sits unreported.

Automated preliminary reporting has therefore become an attractive target for artificial intelligence, because even an imperfect draft report can support triage, prioritise urgent abnormalities, and accelerate communication between radiologists and referring teams when reviewed and corrected. Unlike binary disease classifiers that output only a set of present or absent labels, a draft narrative report provides a first-pass interpretation that a radiologist can quickly verify, amend, and sign. Early image-to-text systems such as TieNet demonstrated that convolutional neural networks could generate free-text impressions from chest radiographs by learning from paired image-report data [1]. Subsequent retrieval-generation models improved fluency by retrieving similar report fragments during decoding [2]. Later work expanded this direction through multimodal curriculum learning, in which models are trained first on simple normal cases then progressively on complex abnormal studies, memory-driven transformers that retain long-range dependencies across multiple findings, and factual consistency objectives that penalise generated statements unsupported by image evidence [3-5]. Together, these advances established that clinically meaningful free text can be generated automatically.

A persistent limitation of many chest X-ray artificial intelligence systems, however, is that they treat interpretation as either disease classification or image captioning, whereas clinical radiology is routinely driven by a request. A referring clinician does not simply ask "what is seen on this image?" but rather submits a free-text request containing symptoms, suspected diagnoses, procedural context, and an explicit or implicit clinical question. A request stating *shortness of breath, query heart failure* requires a very different interpretive emphasis from one that says *rule out pneumothorax after line insertion*, even when the same image contains both mild cardiomegaly and a small apical pneumothorax. In the first case, the radiologist must assess for cardiomegaly, interstitial oedema, pleural effusions, and Kerley B lines; in the second case, attention focuses narrowly on the pleural space along the lateral chest wall, and a tiny pneumothorax that might be dismissed as clinically insignificant in the heart failure query becomes a critical positive finding. Conventional chest X-ray artificial intelligence systems, including most report generators, ignore this crucial framing. Contrastive vision-language models such as ConVIRT, GLoRIA, and ChexZero made significant contributions to aligning chest radiographs with unstructured report text, enabling zero-shot classification and improved visual-language representation learning, but their primary contribution was alignment and recognition rather than request-specific structured reporting [6-8]. Likewise, subsequent report-generation models improved fluency, clinical coverage, and factual consistency, yet they did not consistently encode the referring clinician's free-text question as a first-class input [9-11]. The result is a generated report that may be fluent, sectioned, and superficially complete but fails to answer the explicit clinical query that motivated the study.

Recent multimodal large language models suggest a path toward request-aware interpretation, because they can jointly process images, natural language prompts, and instruction-following behaviour. Models such as XrayGPT, which uses a frozen vision encoder and fine-tuned language model for conversational chest X-ray analysis, and RaDialog, which combines retrieval and dialogue for grounded reporting, demonstrate how chest radiographs can be paired with conversational prompts to produce contextually relevant text [12, 13]. CheXagent, a foundation model for chest radiography with instruction tuning, and MAIRA-2, which introduces grounded generation with spatial localisation of findings, further show that the field is moving toward interactive, prompt-driven interpretation [14, 15]. However, these systems are not necessarily optimised to produce structured preliminary reports that directly answer the clinician's question, separate findings from impressions, flag uncertainty in a format compatible with radiology workflow, and maintain the explicit section headers that referring clinicians rely upon for rapid review. A conceptual framework is therefore needed to translate emerging vision-language model capabilities into a clinically disciplined reporting architecture that treats the free-text request as a central interpretive anchor.

This article proposes a multimodal vision-language model that accepts a current chest X-ray, a free-text radiology request, and optionally prior images or prior reports, then generates a structured preliminary report. The intended output includes an indication restating the request, a technique section, organised findings, an interpretive impression, an explicit answer to the request, and a comparison with prior studies when available. The framework draws on image-text alignment, report generation, retrieval-augmented generation, temporal modelling, and structured evaluation methods developed across recent radiology artificial intelligence literature [16-21]. The article proceeds from background concepts to architecture, multimodal encoding, alignment, structured

generation, safety considerations, workflow integration, and evaluation strategy.

Background

Radiology requests are clinically compact but semantically dense. They typically combine symptoms such as cough or dyspnoea, suspected diagnoses indicated by a question mark before a condition name, procedural context such as post central line placement or post thoracentesis, and a targeted clinical question that may be explicit or implied. Expressions such as *?pneumonia*, *SOB ?CHF*, *rule out effusion*, or *post central line, exclude pneumothorax* direct attention toward particular visual findings and determine which negative observations are clinically important. The same image finding, such as a small pleural effusion, may be reported as clinically insignificant for a request of rule out pneumothorax yet as a key positive finding for a request of query heart failure. Chest X-ray vision-language models must therefore model not only image abnormalities such as airspace opacity, cardiomegaly, pneumothorax, and effusion but also the intent embedded in short, informal, and sometimes ambiguous request text. Prior image-report generation methods demonstrated that chest X-ray findings can be verbalised, and newer multimodal systems show that prompts and clinical language can shape image interpretation, but most existing work treats the prompt as a generic instruction rather than a clinically specific free-text query demanding a direct answer [12, 22, 23].

Structured reporting provides a stable organisation for technique, findings, impression, comparison, and clinically directed conclusions, making reports easier to read, audit, and process with downstream natural language processing tools. In contrast, unstructured generated reports may be fluent but may omit clinically required fields, bury urgent findings within narrative paragraphs, or fail to answer the explicit request. Memory-driven transformer models, cross-modal memory networks, and multi-expert token architectures improved radiology report generation, but their outputs still require constraints to ensure consistent sectioning and clinical completeness [4, 10, 11]. A structured preliminary report decoder can therefore be viewed as a safety layer that converts free-form generation into a reviewable clinical document with predictable organisation, reducing the cognitive load on the reviewing radiologist and ensuring that no required section is omitted.

Multimodal vision-language models generally combine an image encoder, a text encoder or large language model, and an alignment mechanism that allows text tokens to condition visual interpretation. Contrastive pre-training methods such as ConVIRT, GLoRIA, and MedCLIP align radiographs and reports, enabling label-efficient recognition and zero-shot transfer across disease concepts [6, 7, 16]. These approaches learn a shared embedding space where matching image-report pairs are pulled together and non-matching pairs are pushed apart, but they are primarily discriminative rather than generative. Instruction-tuned systems such as XrayGPT and RaDialog extend this paradigm by enabling prompt-driven explanation, question answering, and report-style generation, typically by fine-tuning a large language model with visual tokens projected into its input space [12, 13]. Foundation-model approaches for chest radiography, including CheXagent and grounded systems such as MAIRA-2, further motivate a request-aware model that can combine visual evidence with clinical language while maintaining the ability to follow specific formatting instructions [14, 15].

Chest X-ray report generation has progressed from general encoder-decoder systems toward clinically constrained generation, region-guided explanation, dynamic graph reasoning, and retrieval support. Early encoder-decoder baselines achieved fluency but suffered from factual errors and omission of rare findings. Posterior-prior knowledge distillation used a teacher model based on final reports to guide a student model generating reports, improving disease-specific coverage [9]. Cross-modal memory networks stored visual-textual memory pairs and retrieved relevant memories during decoding, boosting abnormality coverage [10]. Region-guided generation first detected anatomical regions such as the lungs, heart, and pleura, then generated findings per region, improving spatial grounding [19]. Dynamic graph contrastive learning built a graph of visual-semantic relationships, such as connecting cardiomegaly to pulmonary oedema, and updated it dynamically during generation, improving consistency across related findings [24]. More recently, retrieval-augmented generation has offered an additional route by retrieving similar cases, labels, or templates before generating a report, as illustrated by label-boosted retrieval-augmented generation approaches for radiology reporting [21]. These strands support the present framework but must be reorganised around the free-text request as the central interpretive anchor, rather than treating the image alone as the sole input.

Framework Overview

High-level architecture

The proposed system begins with one or more chest radiograph views, which are encoded by a radiographic vision transformer into patch-level visual tokens. The free-text radiology request, available clinical history, and prior report text are encoded by a medical language model into contextual tokens representing symptoms, suspected diagnoses, procedural context, and comparison requirements. Cross-modal attention then permits request tokens to attend to visual regions and prior context before a structured decoder writes the preliminary report. This architecture synthesises principles from image-text embedding networks, contrastive VLMs, report-generation transformers, and multimodal radiology assistants [1, 5, 6, 20].

Figure 1 presents the proposed request-aware multimodal vision-language architecture, showing how chest X-ray images, free-text radiology requests, prior context, cross-modal alignment, structured decoding, safety checks, and radiologist review are organised into a supervised preliminary reporting pathway.

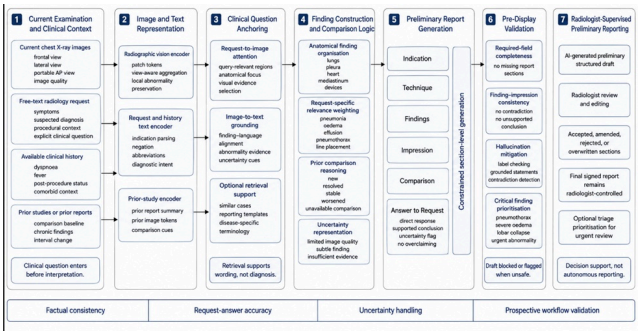

Figure 1. Request-Aware Multimodal Vision-Language Architecture for Structured Preliminary Chest X-Ray Reporting

Core assumptions

The framework assumes access to paired data containing chest X-ray images, radiology requests or indications, final reports, and when available prior studies or report histories. Public datasets such as MIMIC-CXR and related resources are often used for report-generation research, but request text may require careful extraction, harmonisation, and de-identification before model development. Prior work on automatic report generation and evaluation demonstrates that image-report pairs can support multimodal learning, while temporal biomedical VLP shows that prior information can be incorporated when longitudinal structure is available [17, 18, 22]. The model is therefore conceptualised as a data-efficient extension of existing paired image-report learning rather than as an unconstrained free-text generator.

Design principles

The first design principle is structured output: the model should generate a preliminary report with explicit sections rather than a single narrative paragraph. The second principle is direct request-answering, meaning that the impression and answer-to-request field should respond to the referring clinician's stated question whenever the image permits a legitimate conclusion. The third principle is explicit reasoning and safety, drawing on work that improves factual consistency, evaluates report correctness, and uses semantic rewards to discourage clinically unsupported statements [5, 25-27]. These principles position the VLM as a radiologist-support tool rather than an autonomous diagnostic authority.

Multimodal Input Encoding

CXR encoding

The radiographic encoder converts frontal and lateral chest radiographs into visual tokens using patch embeddings, positional encodings, and view-aware aggregation. Multi-view fusion is important because frontal and lateral images can contain complementary evidence, and earlier report-generation work showed that multi-view image fusion can enrich generated radiology descriptions [23]. Region-guided and graph-enhanced models further suggest that visual encoders should preserve local abnormality information rather than collapsing the image too early into a single global vector [19, 24]. In the proposed framework, these visual tokens remain available to cross-modal attention so that the request can selectively query relevant anatomical regions.

Free-text radiology request encoding

The request encoder represents short clinical indications, abbreviations, negated suspicions, and targeted questions as language tokens that guide interpretation. A request such as "fever, cough, ?pneumonia" should activate a different interpretive pathway from "line placement, exclude pneumothorax," even if both are attached to portable chest

radiographs. Biomedical language modelling and report-labeling methods such as CheXbert show that clinically meaningful entities can be extracted from radiology text, while multimodal models such as MedCLIP and XrayGPT demonstrate that medical language can be aligned with visual concepts [12, 16, 25]. The request is therefore encoded not as metadata but as a prompt-like clinical instruction.

Prior study encoding (adaptive)

When prior images or reports are available, the model should adaptively encode them for comparison rather than treating the current image in isolation. Prior radiographs can be represented as additional visual tokens, while prior reports can be summarised into textual tokens describing chronic abnormalities, resolved findings, devices, or baseline cardiopulmonary status. Temporal biomedical vision-language processing indicates that longitudinal structure can be exploited for improved multimodal interpretation, and comparison-aware reporting is essential when the clinical question concerns interval change [18]. The proposed framework therefore includes an adaptive prior-study pathway that is activated only when reliable comparison data exist.

Vision-Language Alignment

Cross-modal attention

Cross-modal attention links the encoded request to image patches so that the model can focus on regions relevant to the question rather than generating a generic report. For example, a request concerning pneumonia should encourage attention to focal air-space opacity, distribution, and associated pleural findings, whereas a pneumothorax request should emphasise pleural lines, lung markings, and support-device context. Global-local alignment methods such as GLoRIA and region-guided report generation show why both whole-image semantics and local evidence are needed for clinically grounded interpretation [7, 19]. In this framework, cross-modal attention acts as the interpretive bridge between the referring clinician's language and the radiographic evidence.

Knowledge retrieval (optional RAG)

Knowledge retrieval can be added when the system benefits from similar solved cases, disease-specific templates, or structured reporting examples. Retrieval-augmented generation should not replace image interpretation, but it can provide priors for report organisation, terminology consistency, and rare-pattern description when appropriately constrained. Label-boosted RAG and other retrieval-oriented radiology report methods illustrate how external case memory can enrich generation while maintaining task relevance [21]. In the proposed model, retrieval is optional, auditable, and subordinated to the current image, request, and available prior comparison.

Structured Report Generation

Hierarchical output schema

The structured decoder should generate the preliminary report as a constrained clinical document rather than as unrestricted prose. A practical schema would include Technique, Findings, Impression, Answer to Request, and Comparison, with the Comparison section activated only when prior imaging or prior reports are available. This design builds on earlier report-generation systems that improved descriptive fluency through retrieval, memory, and transformer decoding, but it adds stronger section-level discipline to make the output easier for radiologists to verify [2, 4, 10]. The key distinction is that the model is not merely asked to "describe the image"; it is required to produce a reviewable draft that separates observations, conclusions, and the direct response to the clinical indication.

Integrating reasoning traces

Before writing the final impression, the model should produce an internal or explicitly reviewable reasoning layer that links visual findings to clinical conclusions. For example, a suspected pneumonia request should connect air-space opacity, distribution, comparison with priors, and alternative explanations before the report states whether the imaging supports infection. Semantic-reward approaches and factual-consistency studies show that radiology generation benefits when clinical correctness is treated as more important than surface fluency [5, 26, 27]. In a preliminary reporting system, reasoning traces should be concise, evidence-linked, and subordinate to the final structured report rather than a substitute for radiologist judgement.

Interpretation of Free-Text Radiology Request

Direct answer tagging

The model should include a dedicated Answer to Request field for radiology requests that contain an implicit or explicit question. This field might state, for example, that there is no radiographic evidence of pneumonia, that pulmonary oedema is likely, or that pneumothorax cannot be excluded because the image is technically limited. Prompt-driven VLMs such as XrayGPT and RaDialog demonstrate why natural-language interaction is useful in radiology, but a clinical reporting system must translate interaction into a stable structured answer [12, 13]. Direct answer tagging therefore turns the free-text request into an organising principle for the report rather than a passive indication line.

Table 1 shows how different free-text chest X-ray requests change the interpretive target, the image evidence that must be prioritised, and the structured reporting obligations of a request-aware VLM.

Table 1. Request-Aware Interpretation Matrix for Chest X-Ray Preliminary Reporting

Radiology request type	Primary clinical intent	Image regions requiring prioritised attention	Finding becc requ criti
“Fever, cough, ?pneumonia”	Determine whether imaging supports infection	Lung zones, focal air-space opacity, distribution, pleural reaction	Consoli patchy c ai broncho assoc effus
“SOB, ?CHF”	Assess for pulmonary oedema or heart failure pattern	Cardiac silhouette, pulmonary vasculature, interstitium, pleural bases	Cardion vasc conge inters oede pleu effus
“Line placement, exclude pneumothorax”	Detect post-procedural complication	Pleural margins, lung apices, device course, mediastinal position	Pneumo malpos lin media sh
“Rule out pleural effusion”	Confirm or exclude pleural fluid	Costophrenic angles, basal pleural spaces, lateral chest wall	Blunt layering unilate bilat effus
“Post-operative hypoxaemia”	Identify acute cardiopulmonary complication	Lungs, pleura, mediastinum, devices, atelectatic regions	Atelec aspira oede pneumo dev compli
“Compare with prior”	Determine interval change	Same anatomical regions across current and prior studies	New, re stable worse abnorr
Technically limited or ambiguous request	Avoid unsupported certainty	Image-quality markers and request-relevant regions	Motion volu rotat incom field, s abnorr

Confidence and uncertainty flagging

A request-aware model must also know when the image does not justify a definitive answer. If the request is unrelated to chest radiography, the image is degraded, the finding is subtle, or prior comparison is essential but unavailable, the model should flag uncertainty instead of inventing a conclusion. Uncertainty-aware medical imaging work, although not specific to CXR reporting, reinforces the importance of representing model confidence when image quality and inference reliability vary [28]. In this framework, uncertainty appears as a structured field that can prompt radiologist review, recommend comparison with prior studies, or suggest correlation with clinical and laboratory findings.

Interpretability and Safety

Visual explanations

Visual explanations should show which image regions contributed to the answer for a specific clinical request. For example, when the request asks about pleural effusion, the explanation should emphasise the costophrenic angles and basal pleural spaces rather than unrelated lung regions. Region-guided report generation and graph-enhanced contrastive methods support this principle because they preserve local abnormality evidence and relate it to generated language [19, 24]. The purpose of these explanations is not to prove that the model is correct, but to make errors more visible during radiologist review.

Hallucination mitigation

Hallucination mitigation should combine label checking, contradiction detection, grounded generation, and conservative uncertainty handling. Automatic labelers such as CheXbert can compare generated statements with extracted radiological observations, while newer evaluation frameworks such as GREEN classify clinically meaningful errors in generated reports [25, 29]. Expert-level self-supervised detection and zero-shot classification methods provide additional checks by testing whether generated findings are supported by independent image-text representations [8, 16]. A safe VLM should therefore be penalised more heavily for unsupported abnormal findings, missed critical findings, and false reassurance than for minor stylistic imperfections.

Structured safety checks

The structured output should undergo automatic validation before it is shown as a preliminary report. Required fields must be present, positive findings should include location and severity where appropriate, and the impression should not contradict the findings or the answer-to-request field. Studies on factual completeness, semantic rewards, and report-generation benchmarking show that clinically useful evaluation must examine consistency, abnormality coverage, and error type rather than relying only on text similarity [5, 26-29]. These safety checks make the model's output more compatible with radiology practice because the draft becomes easier to inspect, correct, and audit.

Clinical Workflow Integration

Real-time preliminary report drafting

In clinical workflow, the VLM should pre-fill a structured report template inside the reporting environment while clearly marking the output as preliminary and AI-generated. The radiologist would then edit, accept, reject, or overwrite each section, preserving human responsibility for the final signed report. Interactive and conversational radiology systems show that VLMs can support image interpretation and report drafting, but deployment requires tight integration with PACS, RIS, and reporting software rather than a detached chatbot interface [13, 20]. The most realistic near-term role is therefore human-in-the-loop drafting, where the model reduces clerical burden without replacing professional interpretation.

Triage and prioritisation

The same framework could support triage by identifying high-confidence critical findings that require faster radiologist attention. Examples include a large pneumothorax after line insertion, severe pulmonary oedema, new lobar collapse, or marked interval worsening when prior comparison is available. Temporal vision-language modelling and grounded radiology report generation are relevant here because triage often depends on both the current image and whether a finding is new, worsening, or clinically expected [15, 18]. Any escalation logic should remain conservative, auditable, and calibrated to local clinical policy so that workflow acceleration does not create alert fatigue.

Evaluation Strategy

Automatic metrics

Evaluation should be structure-aware, request-aware, and clinically grounded. Automatic metrics can assess whether required sections are present, whether findings and impressions are internally consistent, and whether the Answer to Request field matches extracted radiological labels. CheXbert, semantic-reward evaluation, broad benchmarking of CXR report generation, and GREEN-style error notation together provide a foundation for measuring clinical efficacy and clinically meaningful mistakes [25-29]. Text-overlap scores may be reported as secondary descriptive measures, but they should not determine whether the system is safe or useful.

Table 2 provides a structure-aware and request-aware evaluation framework that separates general report quality from the clinically decisive question of whether the model answers the referring clinician's request safely and accurately.

Table 2. Evaluation Framework for a Request-Aware Chest X-Ray Vision-Language Reporting Model

Evaluation domain	Core evaluation question	Suggested assessment method	What success looks like
Section completeness	Does the output contain all required report fields?	Automated schema validation of indication, technique, findings, impression, comparison, and answer-to-request	Every required section is present, correctly labelled, and clinically usable
Request-answer accuracy	Does the model answer the actual clinical question?	Blind radiologist scoring of the answer-to-request field	Response correct, specific, and appropriately limited by available evidence
Visual grounding	Are generated findings supported by image evidence?	Radiologist review with region-level explanation or localisation audit	Positive findings correspond to plausible anatomical evidence
Finding–impression consistency	Does the impression logically follow from findings?	Automated contradiction checks plus expert review	Impression does not introduce unsupported claims or contradict findings
Uncertainty calibration	Does the model avoid overclaiming when evidence is weak?	Review of limited-quality images, subtle findings, and missing priors	Output flags uncertainty, limited image quality, or need for comparison
Comparison accuracy	Does the model correctly use prior studies when available?	Temporal validation using paired current-prior cases	Stable, new, improved, worsened findings are correctly identified
Critical finding sensitivity	Does the model prioritise urgent abnormalities safely?	Enriched test set with pneumothorax, severe oedema, line complications, collapse	Critical abnormalities are detected and flagged for expedited review
Workflow usefulness	Does the draft reduce radiologist burden without reducing safety?	Prospective silent trial followed by monitored deployment study	Radiologist edit efficiently and retain final control

Request-answer accuracy

A central evaluation question is whether the model correctly answers the clinician's request using only what can be concluded from the radiograph and available context. Blind radiologist review should therefore judge whether the generated Answer to Request is correct, overconfident, under-specified, or unsupported by the image. This evaluation is especially important because a report can sound plausible while failing to address the actual reason for imaging, a limitation seen across generic report-generation paradigms [2-4]. Request-answer accuracy should be measured separately from general report quality because it captures the framework's main clinical purpose.

Temporal and cohort validation

The system should be validated retrospectively across institutions, scanners, patient groups, clinical services, and imaging indications before any prospective deployment. Temporal validation is also necessary because request language, imaging protocols, disease prevalence, and reporting conventions can drift over time. Biomedical temporal VLP, grounded report generation, and large-scale report-generation evaluations show that robustness

depends on more than performance on a single static dataset [15, 17, 18, 27]. A staged pathway from silent retrospective testing to monitored prospective use would provide a safer route than immediate autonomous deployment.

Conclusion

A request-aware multimodal vision-language model for chest X-ray interpretation offers a clinically focused alternative to generic image captioning or disease classification. By jointly encoding the radiograph, free-text request, and available prior context, the model can generate structured preliminary reports that are easier to review, edit, and integrate into routine reporting.

The main advantage of the framework is its alignment with how radiologists actually work. It treats the clinical question as central, separates findings from interpretation, supports comparison when prior studies exist, and makes uncertainty visible rather than hiding it inside fluent prose.

Future development should prioritise public datasets enriched with request text, carefully de-identified clinical context, and radiologist-reviewed structured outputs. Prospective clinical validation will be essential to determine whether such systems can improve reporting speed, relevance, safety, and workflow efficiency without weakening professional oversight.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Sep 2025 Revised: 24 Nov 2025 Accepted: 20 Jan 2026
Published online: 20 July 2026

References

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. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2018. p. 9049-58.

Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. Adv Neural Inf Process Syst. 2018;31.

Liu F, Ge S, Wu X. Competence-based multimodal curriculum learning for medical report generation. In: Proc 59th Annu Meet Assoc Comput Linguist Int Joint Conf Nat Lang Process; 2021 Aug. p. 3001-12.

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

Miura Y, Zhang Y, Tsai E, Langlotz C, Jurafsky D. Improving factual completeness and consistency of image-to-text radiology report generation. In: Proc Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol; 2021 Jun. p. 5288-304.

Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Mach Learn Healthc Conf. PMLR; 2022 Dec 31. p. 2-25.

Huang SC, Shen L, Lungren MP, Yeung S. Gloria: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proc IEEE/CVF Int Conf Comput Vis; 2021. p. 3942-51.

Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng. 2022;6(12):1399-406.

Liu F, Wu X, Ge S, Fan W, Zou Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2021. p. 13753-62.

Chen Z, Shen Y, Song Y, Wan X. Cross-modal memory networks for radiology report generation. In: Proc 59th Annu Meet Assoc Comput Linguist Int Joint Conf Nat Lang Process; 2021 Aug. p. 5904-14.

Wang Z, Liu L, Wang L, Zhou L. Metransformer: radiology report generation by transformer with multiple learnable expert tokens. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2023. p. 11558-67.

Thawakar OC, Shaker AM, Mullappilly SS, Cholakkal H, Anwer RM, Khan S, et al. Xraygpt: chest radiographs summarization using large medical vision-language models. In: Proc 23rd Workshop Biomed Nat Lang Process; 2024 Aug. p. 440-48.

Pellegrini C, Özsoy E, Busam B, Navab N, Keicher M. Radialog: a large vision-language model for radiology report generation and conversational assistance. arXiv. 2023;arXiv:2311.18681.

Chen Z, Varma M, Xu J, Paschali M, Van Veen D, Johnston A, et al. A vision-language foundation model to enhance efficiency of chest X-ray interpretation. arXiv. 2024;arXiv:2401.12208.

Bannur S, Bouzid K, Castro DC, Schwaighofer A, Thieme A, Bond-Taylor S, et al. Maira-2: grounded radiology report generation. arXiv. 2024;arXiv:2406.04449.

Wang Z, Wu Z, Agarwal D, Sun J. Medclip: contrastive learning from unpaired medical images and text. In: Proc Conf Empir Methods Nat Lang Process; 2022 Dec. p. 3876-87.

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. In: Eur Conf Comput Vis. Cham: Springer; 2022. p. 1-21.

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. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2023. p. 15016-27.

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

Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. Eur Radiol. 2025;35(7):4374-86.

Song S, Subramanyam A, Madejski I, Grossman RL. Lab-rag: label boosted retrieval augmented generation for radiology report generation. arXiv. 2024;arXiv:2411.16523.

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

Yuan J, Liao H, Luo R, Luo J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Int Conf Med Image Comput Comput Assist Interv. Cham: Springer; 2019. p. 721-29.

Li M, Lin B, Chen Z, Lin H, Liang X, Chang X. Dynamic graph enhanced contrastive learning for chest X-ray report generation. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2023. p. 3334-43.

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. In: Proc Conf Empir Methods Nat Lang Process; 2020. p. 1500-19.

Delbrouck JB, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz C. Improving the factual correctness of radiology report generation with semantic rewards. In: Findings Assoc Comput Linguist EMNLP; 2022. p. 4348-60.

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).

Tanno R, Worrall DE, Ghosh A, Kaden E, Sotiropoulos SN, Criminisi A, et al. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In: Int Conf Med Image Comput Comput Assist Interv. Cham: Springer; 2017. p. 611-19.

Ostmeier S, Xu J, Chen Z, Varma M, Blankemeier L, Bluethgen C, et al. Green: generative radiology report evaluation and error notation. In: Findings Assoc Comput Linguist EMNLP; 2024. p. 374-90.