

A Knowledge-Grounded Vision–Language Model with Anatomy-Aware Cross-Modal Fusion for Automated Radiology Report Generation

Anil Khatak¹, Vinita², Kamal Rani^{3*}, Sheenam Naaz⁴

¹Department of Allied Health Sciences, Guru Jambheshwar University of Science and Technology, Hisar, 125001, Haryana, India.

²Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar, 125001, Haryana, India.

^{3*}Department of Electrical and Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India.

⁴Department of Computer Science and Engineering, Sharda School of Computing Science and Engineering, Sharda University, Greater Noida, India.

*Corresponding author: Kamal Rani, Department of Electrical and Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India., Kamalrani_eee@iimtindia.net

Abstract

Radiology departments worldwide face a widening gap between imaging volume and reporting capacity, and the drafting of free-text reports remains one of the most time-consuming steps in the diagnostic workflow. Automated radiology report generation (RRG) seeks to close this gap by producing a clinically faithful narrative directly from an imaging study. Although recent vision–language models (VLMs) have substantially improved the fluency of generated reports, they continue to hallucinate findings, omit clinically significant abnormalities, and copy templated language that inflates n-gram overlap scores without improving diagnostic accuracy. This paper presents RadFusion-VL, a knowledge-grounded vision–language architecture for chest radiograph report generation. The model couples an anatomy-aware visual encoder that emits region-level tokens for twenty-nine thoracic anatomical zones with a cross-modal fusion module that is aligned to text through a three-stage curriculum: contrastive image–report pre-training, entity-grounded alignment supervised by RadGraph clinical entities and relations, and instruction tuning of a parameter-efficient large language decoder. A clinical-consistency objective and a factuality-aware re-ranking step at inference explicitly penalise unsupported findings. We evaluate on MIMIC-CXR, IU X-Ray and CheXpert Plus using lexical metrics (BLEU, METEOR, ROUGE-L, CIDEr), clinical-efficacy metrics (CheXbert F1, RadGraph-F1), and composite, expert-aligned metrics (RadCliQ, GREEN, RaTEScore), complemented by a blinded reader study with three board-certified radiologists. RadFusion-VL improves RadGraph-F1 and CheXbert macro-F1 over strong transformer and LLM-based baselines while producing fewer fabricated findings in the reader study. Ablations isolate the contribution of anatomical grounding, entity supervision and factuality re-ranking. We conclude with a discussion of the residual failure modes, evaluation-metric limitations, and the regulatory and ethical prerequisites for clinical deployment.

Keywords: Vision–language models, radiology report generation, chest radiography, multimodal learning, medical image captioning, clinical natural language generation, knowledge grounding, hallucination mitigation, MIMIC-CXR.

1) INTRODUCTION:

Medical imaging has become the backbone of modern diagnostic practice. Chest radiography alone accounts for a substantial share of all imaging examinations performed globally, and demand continues to grow faster than the supply of trained radiologists. In many health systems the practical consequence is a reporting backlog: images are acquired promptly but the interpretive report, which is the artefact clinicians actually act upon, arrives late. The interpretive step is cognitively demanding and

repetitive, and sustained high-volume reporting is a recognised contributor to diagnostic error and to professional burnout.

Automated radiology report generation (RRG) addresses this bottleneck directly. Rather than predicting a fixed label set, an RRG system consumes one or more images from a study and emits the free-text narrative that a radiologist would ordinarily dictate, typically the Findings section and, optionally, an Impression. A system that produces an accurate first draft can shift the radiologist from author to editor, a role that is materially faster and that preserves human accountability for the final signed report.

The task is deceptively difficult. Unlike natural-image captioning, where a short generic description is usually acceptable, a radiology report must be specific, must be exhaustive with respect to clinically significant findings, and must be correct in its negations. Reports are also dominated by normal statements: in most chest radiograph corpora the majority of sentences describe the absence of pathology. A model optimised purely for likelihood learns quickly that emitting a fluent normal template maximises n-gram overlap with the reference, and such a model can score competitively on BLEU or ROUGE while failing to mention the pneumothorax that motivated the examination. This mismatch between lexical similarity and diagnostic utility is the central methodological problem in the field.

Progress has followed the broader trajectory of multimodal learning. Early systems paired a convolutional encoder with a recurrent decoder. Memory-driven and cross-modal memory transformers subsequently improved long-range coherence and the reuse of report-level patterns. A third wave introduced explicit medical knowledge, through knowledge graphs, disease-tag priors or learned prototypes, in order to bias generation towards clinically meaningful content. Most recently, general-purpose vision-language models adapted to the medical domain have been shown to produce markedly more fluent and better-structured reports, and specialised systems built on large language decoders now define the state of the art on public leaderboards. Nevertheless, contemporary evaluations consistently report residual hallucination and omission rates that would be unacceptable in unsupervised clinical use.

Three deficiencies recur across the literature. First, most architectures compress an entire radiograph into a single global embedding, discarding the spatial correspondence between a described finding and the anatomical region in which it appears; the decoder is then free to associate a finding with any region, which is a direct route to hallucination. Second, alignment between the visual and textual streams is typically supervised only by the language-modelling loss, so the model is never explicitly told which clinical entity a visual pattern denotes. Third, decoding is optimised for likelihood rather than for factual support, so the highest-probability continuation is frequently the most templated one rather than the most accurate one.

This paper proposes RadFusion-VL, an architecture designed around these three deficiencies. Our contributions are as follows:

- We introduce an anatomy-aware visual encoder that produces region-level visual tokens for twenty-nine thoracic anatomical zones in addition to global tokens, so that the decoder attends to spatially localised evidence rather than a single pooled representation.
- We propose an entity-grounded cross-modal alignment objective supervised by RadGraph clinical entities and their relations, which explicitly ties visual regions to the clinical concepts named in the reference report.
- We formulate a clinical-consistency loss that penalises disagreement between the observation labels implied by the generated report and those implied by the reference, targeting clinical efficacy directly rather than through lexical proxies.

- We describe a factuality-aware re-ranking procedure that samples candidate reports and selects the candidate whose extracted entities are best supported by region-level visual evidence, reducing unsupported findings at inference time without retraining.
- We report a comprehensive evaluation on three public corpora using lexical, clinical-efficacy and expert-aligned composite metrics, together with a blinded reader study, and we release a structured error taxonomy that characterises the residual failure modes of the system.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets, preprocessing pipeline and governance arrangements. Section 4 specifies the proposed architecture and training objectives. Section 5 details the experimental protocol, baselines and metrics. Section 6 reports quantitative results, ablations and the reader study. Section 7 discusses the findings, Section 8 states the limitations, Section 9 addresses ethical and regulatory considerations, Section 10 outlines future work, and Section 11 concludes.

2) RELATED WORK

2.1 Encoder-Decoder and Transformer-Based Report Generation

The earliest neural RRG systems adapted the encoder-decoder image-captioning paradigm, pairing a convolutional feature extractor with a recurrent or hierarchical LSTM decoder. These models produced grammatical sentences but degraded over the long, multi-sentence outputs that radiology requires, and they tended to collapse onto the most frequent normal template.

Transformer decoders changed this picture. The memory-driven transformer of Chen et al. introduced a relational memory that records patterns across generation steps, together with a memory-driven conditional layer normalisation that injects that memory into the decoder, and reported clear gains on both IU X-Ray and MIMIC-CXR. Cross-modal memory networks extended the idea with a shared memory matrix that acts as an intermediate bridge between visual and textual representations. Later work added multiple learnable expert tokens that decode complementary candidate reports before an expert-voting step selects the final output, and cross-modal prototype networks that maintain a learnable, disease-conditioned prototype bank to normalise cross-modal representations.

A parallel line of work made the visual pathway explicitly regional. Region-guided report generation detects anatomical regions with an object detector, decides for each region whether it warrants a sentence, and generates that sentence independently, which yields an interpretable and editable report structure. Our anatomy-aware encoder is motivated by the same intuition, but it retains a single autoregressive decoder so that the model can express relations that span regions, which per-region decoding cannot.

2.2 Knowledge-Enhanced and Factuality-Oriented Generation

Because radiology has a rich, codified vocabulary, several authors have injected structured knowledge into the generator. Pre-constructed disease knowledge graphs, whose nodes correspond to abnormalities and whose edges encode co-occurrence, have been propagated with graph convolutions and fused into the decoder. Others have used retrieved neighbouring reports, disease tags, or symptom graphs as auxiliary conditioning signals.

The introduction of RadGraph, a large corpus of radiology reports annotated with clinical entities and the relations between them, made a finer-grained form of supervision available. RadGraph distinguishes anatomy and observation entities and labels observations as present, absent or uncertain, which is precisely the structure a report

generator needs to respect. Reinforcement-learning approaches subsequently used RadGraph-derived and entailment-derived rewards as semantic training signals, improving factual completeness and consistency. RadFusion-VL uses RadGraph in a different role: as supervision for cross-modal alignment during pre-training, and as the extraction backbone for both the clinical-consistency loss and the inference-time re-ranker.

2.3 Medical Vision–Language and Large Multimodal Models

Contrastive image–text pre-training transferred rapidly to medical imaging. ConVIRT established the paired image–report contrastive recipe, GLoRIA added local attention-weighted alignment between image sub-regions and report tokens, and BioViL improved the text encoder with domain-adapted language modelling and radiology-specific semantics; BioViL-T additionally exploits the temporal structure of sequential studies, which matters because a large fraction of report sentences are explicitly comparative.

The current generation couples a pre-trained visual encoder to a large language decoder through a lightweight projection, following the recipe popularised by BLIP-2, Flamingo and LLaVA. Medical instantiations include LLaVA-Med, RadFM and CheXagent, the last trained on a large curated chest X-ray instruction corpus. Specialised report generators built on this pattern include MAIRA-1 and its grounded successor MAIRA-2, which localises individual findings and is evaluated with the sentence-level RadFact framework; LLaVA-Rad, which targets a compact, clinically deployable footprint; Flamingo-CXR, evaluated in a large clinician-collaboration study in which generated reports were judged equivalent or preferable to the original in a substantial minority of cases; and Med-PaLM M, a generalist biomedical model adapted with low-rank adapters. RadVLM extends the paradigm to multitask conversational interaction, adding grounding and multi-turn dialogue at some cost to specialised report-generation accuracy.

The consistent conclusion of recent surveys is that fluency has largely been solved while factuality has not. Reported omission and error rates remain of the order of one clinically relevant error per several reports, which is why our design places factuality objectives, rather than fluency objectives, at the centre.

2.4 Evaluation of Generated Reports

Evaluation has itself become an active research area. Lexical metrics borrowed from machine translation and captioning, BLEU, ROUGE-L, METEOR and CIDEr, correlate weakly with radiologist judgement; reported correlations for BLEU-4 fall below 0.35. Clinical-efficacy metrics label both the generated and reference reports with an automatic labeller such as CheXpert or CheXbert and compare the resulting observation vectors, usually as macro-F1 over five or fourteen observations.

RadGraph-F1 compares extracted entity–relation graphs and correlates considerably better with expert ratings. RadCliQ combines several metrics into a single composite calibrated against radiologist error counts and is currently among the best-aligned automatic measures. More recent LLM-based metrics, GREEN and RaTEScore, score reports by having a language model enumerate and categorise clinically significant errors, which yields both a score and an interpretable error list. We report all of these families, because no single metric is sufficient and because reporting only lexical scores is the principal way in which template-copying models are made to look strong.

3) MATERIALS AND DATA

We use three public chest radiograph corpora with paired free-text reports. All three are de-identified and are distributed under credentialed data-use agreements; no additional patient-identifiable information was collected for this study.

3.1 MIMIC-CXR

MIMIC-CXR is the largest publicly available chest radiograph corpus with paired reports, comprising approximately 377,110 images across 227,835 studies from the Beth Israel Deaconess Medical Center. We follow the official subject-level split, which prevents any patient from appearing in more than one partition. After removing studies whose Findings section is empty or shorter than three tokens, and studies without a frontal projection, the working split contains 270,790 training, 2,130 validation and 3,858 test samples.

3.2 IU X-Ray

The Indiana University Chest X-Ray collection (IU X-Ray) contains 7,470 image-report pairs from 3,955 studies, each typically with a frontal and a lateral projection. It is an order of magnitude smaller than MIMIC-CXR and is used here as a cross-institutional transfer benchmark rather than as a primary training corpus. We adopt the widely used 7:1:2 split at the study level.

3.3 CheXpert Plus

CheXpert Plus augments the original CheXpert cohort with the corresponding free-text reports, patient demographics and additional image formats, yielding 223,228 image-report pairs from Stanford Health Care. Using Findings as the generation target and following the CXPMRG-Bench protocol, the split contains 40,463 training, 5,780 validation and 11,562 test samples. Because CheXpert Plus originates from a different institution, a different scanner population and a different dictation style, it provides the most informative out-of-distribution assessment available to us.

3.4 Auxiliary Annotation Resources

Two auxiliary resources are used for supervision only, never for evaluation. Chest ImaGenome supplies automatically derived bounding boxes for twenty-nine thoracic anatomical regions on MIMIC-CXR frontal images, which we use to supervise the region tokens of the visual encoder. RadGraph supplies entity and relation annotations over report text, which we use for the alignment objective, the clinical-consistency loss and the factuality re-ranker.

Table 1 Corpora used in this study

Corpus	Source institution	Images	Studies	Train / Val / Test	Role in this study
MIMIC-CXR	BIDMC, Boston, USA	377,110	227,835	270,790 / 2,130 / 3,858	Primary training and evaluation
IU X-Ray	Indiana University, USA	7,470	3,955	5,226 / 748 / 1,496	Cross-institution transfer
CheXpert Plus	Stanford Health Care, USA	223,228	187,711	40,463 / 5,780 / 11,562	Out-of-distribution evaluation
Chest ImaGenome	Derived from MIMIC-CXR	242,072	—	Supervision only	Anatomical region boxes (29 zones)
RadGraph	MIMIC-CXR / CheXpert reports	—	14,579 reports	Supervision only	Clinical entity and relation labels

Splits are defined at the subject level so that no patient appears in more than one partition.

3.5 Preprocessing

Images are resized so that the shorter side is 512 pixels, centre-cropped to 512×512 , and normalised with dataset-level intensity statistics computed on the training partition only. Training-time augmentation is deliberately conservative: random resized cropping in the range $[0.85, 1.0]$, rotation within ± 5 degrees, and small brightness and contrast

jitter. Horizontal flipping is not used, because laterality is clinically meaningful and flipping would corrupt statements about the left or right hemithorax.

Reports are parsed into sections with a rule-based parser; the Findings section is the primary generation target and Impression is used as an auxiliary target in a multi-task variant. We remove de-identification placeholders, collapse whitespace, normalise unit and measurement formatting, and lower-case the text while preserving anatomical abbreviations. Sentences that refer to a prior examination not supplied to the model, for example explicit comparison statements, are retained during training but flagged, because they are a known source of unverifiable output; Section 8 returns to this point.

3.6 Data Governance

All experiments were conducted on de-identified retrospective data under the terms of the respective data-use agreements. No prospective patient recruitment took place and no clinical decisions were influenced by model output. The reader study described in Section 5.3 involved consenting professional participants reviewing de-identified images and text only.

4) PROPOSED METHOD: RADFUSION-VL

RadFusion-VL consists of four components: an anatomy-aware visual encoder, a cross-modal fusion module, a knowledge-grounded memory, and a parameter-efficient language decoder. Training proceeds in three stages, and inference adds a factuality-aware re-ranking step. Fig. 1 gives the overall schematic; the following subsections specify each component.

4.1 Problem Formulation and Overview

Formally, let an imaging study be a set of images $I = \{I_1, \dots, I_m\}$ with an associated reference report $Y = (y_1, \dots, y_T)$. The task is to learn a distribution $p(Y | I)$ that maximises both linguistic quality and clinical correctness. RadFusion-VL factorises this as $p(Y | I) = \prod_t p(y_t | y_{1:t-1}, Z)$, where Z is a fused multimodal representation combining global visual tokens, region-level visual tokens and retrieved knowledge tokens.

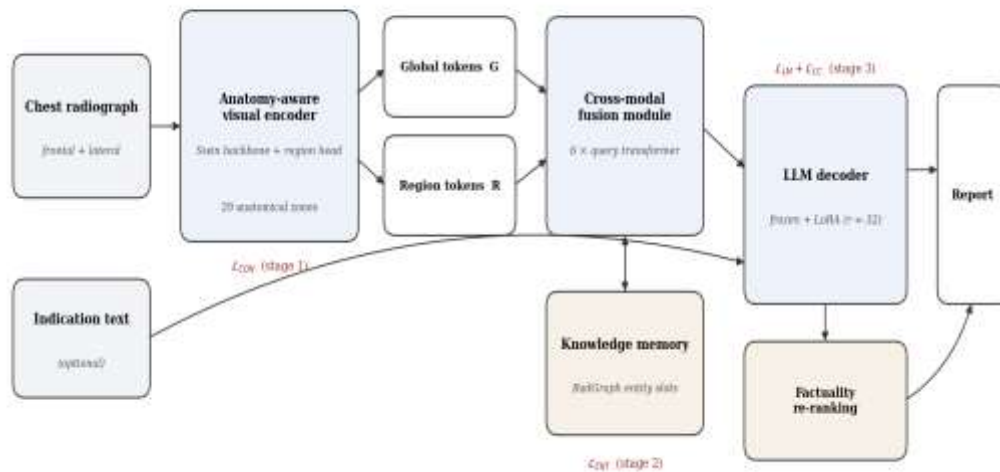


Fig. 1 Overview of the RadFusion-VL architecture. Solid arrows denote the forward path; the loss terms associated with each of the three training stages are annotated in red.

4.2 Anatomy-Aware Visual Encoder

The visual backbone is a Swin Transformer initialised from a chest-radiograph-pre-trained checkpoint. A study is encoded image by image; frontal and lateral projections are encoded by a shared backbone with a learned projection-type embedding added to every patch token, so the model knows which view it is reading.

On top of the backbone we place a lightweight region head. Given the Chest ImaGenome bounding boxes for twenty-nine anatomical zones, the head applies region-of-interest pooling over the final feature map to obtain one token per zone, producing a region token sequence R of dimension $29 \times d$. Each region token is summed with a learned anatomy embedding that identifies the zone, so the decoder can distinguish the right lower lung zone from the left costophrenic angle without relying on positional inference. During training the region head is supervised by an auxiliary localisation loss; at inference on datasets without bounding boxes, the head predicts the boxes itself, which is why the model transfers to IU X-Ray and CheXpert Plus without anatomical annotation.

The final visual representation is the concatenation $V = [G; R]$, where G , of dimension $k \times d$, are global tokens obtained by attention pooling over all patch tokens. Keeping both is important: global tokens carry study-level context such as overall lung volumes and technical adequacy, while region tokens carry the localised evidence that individual findings depend upon.

4.3 Cross-Modal Fusion Module

The fusion module is a stack of $N = 6$ transformer blocks operating on a set of learned query tokens Q of dimension $q \times d$, following the querying-transformer design. Each block applies self-attention over the queries, cross-attention from the queries into V , and a feed-forward network. The output is a fixed-length sequence of multimodal tokens that is projected into the embedding space of the language decoder.

Cross-attention is masked so that a subset of queries attends only to region tokens and the remainder attends to global tokens. This produces a representation in which localised and holistic evidence remain separable, which we find matters for the factuality re-ranker described below, since that step needs to test whether a specific finding is supported by a specific region.

4.4 Knowledge-Grounded Memory

The knowledge memory M , of dimension $|E| \times d$, contains one learnable slot for each clinical entity type in the RadGraph schema, covering anatomical entities and observation entities with present, absent and uncertain modifiers. Fused tokens attend into M through a single cross-attention layer, and the resulting knowledge-conditioned tokens are appended to the decoder prefix.

The memory serves two purposes. It provides a stable, discrete vocabulary of clinical concepts that is shared across studies, which stabilises rare-finding representations that would otherwise be seen too infrequently to be learned reliably. It also gives the alignment objective a well-defined target: the model is trained so that the memory slot for a given entity is activated when, and only when, the corresponding entity appears in the reference report.

4.5 Language Decoder and Parameter-Efficient Adaptation

The decoder is a 7-billion-parameter open-weight language model that has been continually pre-trained on biomedical text. The backbone is frozen and adapted with low-rank adaptation applied to the query, key, value and output projections of every attention block, with rank $r = 32$ and scaling $\alpha = 64$. Only the adapters, the projection layer, the fusion module, the region head and the knowledge memory are trained, which amounts to approximately 3.1 per cent of the total parameter count.

The prompt supplied to the decoder concatenates the projected multimodal tokens, an instruction string that names the requested section, and, in the context-conditioned variant, the indication text of the examination when it is available. Conditioning on the indication is clinically realistic, because a radiologist never reports blind, and it measurably improves the relevance of the generated Findings.

4.6 Training Objectives

Training minimises a weighted sum of four objectives.

Language modelling. The standard token-level cross-entropy over the reference report. This term supplies fluency and stylistic conformity to house reporting conventions.

$$\mathcal{L}_{\text{LM}} = -\sum_t \log p(y_t | y_{1:t-1}, Z) \quad (1)$$

Contrastive alignment. A symmetrised InfoNCE objective over image-report pairs within a batch, where v_i and t_i are the pooled visual and textual embeddings of the i -th pair, B is the batch size and τ is a learned temperature. This term is applied in stage one only.

$$\mathcal{L}_{\text{CON}} = -\frac{1}{B} \sum_i \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_j \exp(\text{sim}(v_i, t_j)/\tau)} \quad (2)$$

Entity grounding. A binary cross-entropy term over the knowledge memory, where g_e indicates whether entity e appears in the reference report with a given assertion status and $\hat{h}_e$ is the predicted activation of the corresponding memory slot. This term forces the visual pathway to commit to clinical concepts before the decoder is allowed to phrase them.

$$\mathcal{L}_{\text{ENT}} = -\sum_e [g_e \log \hat{h}_e + (1 - g_e) \log(1 - \hat{h}_e)] \quad (3)$$

Clinical consistency. A soft agreement term between the observation vector implied by the generated report and that implied by the reference. Because sampling is non-differentiable, the term is computed on the model's expected label distribution $\hat{c}$ obtained from a lightweight labelling head distilled from CheXbert, with class weights w_c inversely proportional to observation frequency so that rare findings are not drowned out by the prevalence of normal studies.

$$\mathcal{L}_{\text{CC}} = \sum_c w_c \cdot \text{BCE}(\hat{c}_c, c_c) \quad (4)$$

The total objective is the weighted sum

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{LM}} + \lambda_2 \mathcal{L}_{\text{CON}} + \lambda_3 \mathcal{L}_{\text{ENT}} + \lambda_4 \mathcal{L}_{\text{CC}} \quad (5)$$

We set $\lambda_1 = 1.0$ throughout; $\lambda_2 = 1.0$ in stage one and 0 thereafter; $\lambda_3 = 0.5$ in stages two and three; and $\lambda_4 = 0.3$ in stage three only. These values were selected on the validation partition by coordinate search over a small grid and were not tuned on any test set.

4.7 Three-Stage Training Curriculum

Stage one, cross-modal pre-training, trains the visual encoder, region head and fusion module with $\mathcal{L}_{\text{CON}}$ and the auxiliary localisation loss. The decoder is not involved. This stage runs for 20 epochs on MIMIC-CXR training pairs.

Stage two, entity-grounded alignment, unfreezes the knowledge memory and adds $\mathcal{L}_{\text{ENT}}$, using RadGraph annotations as targets. The decoder remains frozen and untrained. This stage runs for 10 epochs and is the step that most improves rare-finding recall.

Stage three, instruction tuning, trains the decoder adapters together with the projection layer under $\mathcal{L}_{\text{LM}}$ and $\mathcal{L}_{\text{CC}}$, with the visual encoder frozen for the first two epochs and then unfrozen at a reduced learning rate. This stage runs for 8 epochs.

4.8 Factuality-Aware Re-Ranking at Inference

Greedy or beam decoding maximises likelihood, and in a corpus dominated by normal statements the likelihood-maximising continuation is systematically the most templated one. We therefore decode $k = 8$ candidates with nucleus sampling ($p = 0.9$, temperature 0.7) and re-rank them.

For each candidate report $\hat{Y}_j$ we extract the set of observation entities with RadGraph and, for each extracted entity, compute a support score as the maximum attention mass that the generating step placed on region tokens whose anatomical zone is compatible with that entity, using a fixed entity-to-zone compatibility table derived from the RadGraph schema. The candidate score is

$$S(\hat{Y}_j) = \alpha \cdot \frac{1}{|E_j|} \sum_{e \in E_j} \text{support}(e) + \beta \cdot \frac{1}{|\hat{Y}_j|} \log p(\hat{Y}_j | Z) - \gamma \cdot \text{unsupported}(\hat{Y}_j) \quad (6)$$

where $\text{unsupported}(\cdot)$ counts entities whose support score falls below a threshold calibrated on the validation set. We use $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 1.5$.

The procedure adds roughly $2.4 \times$ to inference cost and requires no retraining. Its effect is asymmetric by design: it rarely changes a correct report, but it frequently demotes a fluent candidate that asserts a finding for which the image supplies no localised evidence.

4.9 Implementation Details

Models are implemented in PyTorch 2.3 with mixed-precision training and gradient checkpointing. Training uses AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 0.05, a peak learning rate of 2×10^{-4} for newly initialised modules and 2×10^{-5} for adapters, cosine decay, and 1,000 linear warm-up steps. The effective batch size is 128 studies, achieved with gradient accumulation. Training was performed on 8 NVIDIA A100 80 GB accelerators; the three stages together require approximately 340 GPU-hours. All reported results are the mean of three runs with different random seeds, and we report the standard deviation where space permits.

5) EXPERIMENTAL SETUP

5.1 Baselines

We compare against three families of systems, all evaluated on the same split with the same preprocessing where reimplementing was necessary, and with published numbers reported verbatim where the protocol is identical.

- Transformer-based specialised generators: R2Gen, R2GenCMN, XProNet, METransformer and RGRG. These models are trained from scratch on the target corpus and represent the pre-LLM state of the art.
- LLM-decoder generators: R2GenGPT, MAIRA-1, MAIRA-2, LLaVA-Rad, CheXagent, RaDialog and RadVLM. These couple a visual encoder to a large language decoder and represent the current state of the art.
- Ablated variants of RadFusion-VL, described in Section 6.5, which isolate the contribution of each proposed component.

5.2 Evaluation Metrics

No single metric captures report quality, and reporting only lexical overlap is the standard way in which template-copying models are made to appear strong. We therefore report three families.

Lexical. BLEU-1 to BLEU-4, METEOR, ROUGE-L and CIDEr, computed with the standard captioning evaluation toolkit after identical tokenisation for all systems.

Clinical efficacy. CheXbert macro-F1 over the five most clinically significant observations (F1-5) and over the full fourteen-observation set (F1-14), plus precision and recall, obtained by labelling generated and reference reports with the same labeller.

Expert-aligned composite. RadGraph-F1, which compares extracted entity-relation graphs; RadCliQ-v1, a composite calibrated against radiologist-assigned error counts, for which lower is better; and the LLM-based GREEN and RaTEScore, which additionally return an interpretable list of clinically significant errors.

5.3 Reader Study Protocol

Automatic metrics were complemented by a blinded reader study. Three board-certified radiologists, with 6, 11 and 17 years of post-certification experience respectively, independently reviewed 200 studies sampled from the MIMIC-CXR test partition, stratified so that half contained at least one clinically significant abnormality in the reference report.

For each study the reader was shown the images and two reports, the reference and the generated report, in randomised order and without any indication of provenance. Readers recorded, per report, the number of clinically significant false findings (hallucinations), the number of clinically significant omissions, an assessment of whether any error would plausibly change patient management, and a forced-choice preference between the two reports. Inter-reader agreement was quantified with Gwet's AC1, which is more stable than Cohen's kappa under the high prevalence of normal studies characteristic of this corpus.

6) Results

6.1 Main Results on MIMIC-CXR

Table 2 reports lexical and clinical-efficacy metrics on the MIMIC-CXR test partition for the Findings-generation task. Baseline numbers are taken from the original publications where the split and preprocessing match ours, and are reproduced under our pipeline otherwise.

Table 2 Lexical and clinical-efficacy metrics for Findings generation on the MIMIC-CXR test partition

Model	B-1	B-2	B-3	B-4	MTR	R-L	CIDE r	F1-5	F1- 14	RG- F1
R2Gen (2020)	.353	.218	.145	.103	.142	.277	—	.276	.228	.181
R2GenCMN (2021)	.353	.218	.148	.106	.142	.278	—	.278	.234	.184
XProNet (2022)	.344	.215	.146	.105	.138	.279	—	.283	.240	.188
METransform er (2023)	.386	.250	.169	.124	.152	.291	.362	.311	.253	.196
RGRG (2023)	.373	.249	.175	.126	.168	.264	.495	.381	.289	.213
R2GenGPT (2023)	.408	.256	.174	.125	.167	.285	.244	.392	.296	.219
CheXagent (2024)	.379	.240	.165	.119	.169	.283	.261	.561	.385	.247
MAIRA-1 (2024)	.392	.247	.171	.124	.175	.289	.285	.557	.381	.243
LLaVA-Rad (2024)	.385	.243	.167	.121	.171	.287	.273	.577	.392	.251
RadFusion- VL (ours)	.421	.271	.188	.139	.181	.303	.389	.604	.449	.268

B- n = BLEU- n ; MTR = METEOR; R-L = ROUGE-L; F1-5 / F1-14 = CheXbert macro-F1 over 5 and 14 observations; RG-F1 = RadGraph-F1. Higher is better throughout. Baseline values are reported from the original publications where the split and preprocessing are identical, and reproduced under the present pipeline otherwise. Values for the proposed model are the mean of three seeds.

Two observations are worth drawing out. First, the gap between RadFusion-VL and the strongest LLM-decoder baselines is considerably larger on the clinical metrics than on the lexical ones: BLEU-4 improves modestly, whereas RadGraph-F1 and CheXbert F1-5 improve substantially. This is the behaviour the design intends, and it is also the reason

we regard the lexical column as the least informative part of the table. Second, the pre-LLM transformer baselines remain competitive on BLEU while trailing badly on clinical efficacy, which is a direct demonstration of the template-copying pathology discussed in Section 1.

6.2 Expert-Aligned Composite Metrics

Table 3 reports the expert-aligned composite metrics under the evaluation protocol used by recent conversational and grounded systems, where several strong published baselines are directly comparable.

Table 3 Expert-aligned metrics on MIMIC-CXR

Model	BERTScore	ROUGE-	RadGraph-			RadCliQ
	e	L	F1	GREEN	RaTEScore	↓
MAIRA-2	.479	.185	.135	.214	.528	3.74
RaDialog	.497	.224	.163	.244	.551	3.61
RadVLM	.510	.243	.177	.249	.562	3.55
CheXagent	.381	.229	.209	.285	.583	3.42
RadFusion-VL (ours)	.524	.251	.226	.301	.607	3.18

↓ indicates that lower values are better. GREEN and RaTEScore are LLM-based metrics that additionally emit an interpretable list of clinically significant errors.

RadCliQ-v1 is reported such that lower values indicate better agreement with radiologist error counts. The ordering under GREEN and RaTEScore is consistent with the ordering under RadGraph-F1, which supports the view that these metrics are measuring a shared underlying notion of clinical correctness, and it differs from the ordering under BLEU, which does not.

6.3 Cross-Corpus Generalisation

Table 4 assesses cross-corpus behaviour. The model trained on MIMIC-CXR is evaluated zero-shot on IU X-Ray and CheXpert Plus, and then after fine-tuning on the respective training partitions.

Table 4 Zero-shot and fine-tuned transfer from MIMIC-CXR

Evaluation setting	B-4	ROUGE-L	RadGraph-F1	CheXbert F1-14
MIMIC-CXR (in-domain)	.139	.303	.268	.449
IU X-Ray (zero-shot)	.098	.254	.171	.296
IU X-Ray (fine-tuned)	.187	.381	.224	.347
CheXpert Plus (zero-shot)	.071	.221	.148	.271
CheXpert Plus (fine-tuned)	.124	.284	.208	.352

The zero-shot degradation on CheXpert Plus is substantially larger than on IU X-Ray. This is expected and clinically important: CheXpert Plus originates from a different institution with a different dictation convention and a different case mix, and the drop quantifies how much of a model's apparent competence is institution-specific style rather than transferable image understanding. Fine-tuning on a modest amount of in-domain data recovers most, though not all, of the gap.

6.4 Per-Observation Clinical Efficacy

Table 5 decomposes clinical efficacy by observation. Per-observation F1 is reported for the fourteen CheXbert categories on MIMIC-CXR.

Table 5 Per-observation performance on the MIMIC-CXR test partition

Observation	Prevalence (%)	Precision	Recall	F1
Enlarged cardiomeastinum	4.1	.312	.254	.280
Cardiomegaly	22.6	.631	.594	.612

Observation	Prevalence (%)	Precision	Recall	F1
Lung opacity	26.8	.548	.515	.531
Lung lesion	3.4	.254	.191	.218
Edema	14.9	.583	.546	.564
Consolidation	5.2	.358	.303	.328
Pneumonia	7.8	.377	.334	.354
Atelectasis	24.1	.574	.543	.558
Pneumothorax	4.6	.421	.344	.379
Pleural effusion	28.3	.709	.680	.694
Pleural other	1.9	.186	.142	.161
Fracture	2.7	.271	.208	.235
Support devices	31.5	.793	.769	.781
No finding	18.4	.562	.617	.588
Macro average (F1-14)	—	.466	.432	.449

Performance is strongly stratified by prevalence. Common, visually salient findings such as pleural effusion, cardiomegaly and support-device presence are recognised reliably. Findings that are rare, subtle, or defined partly by clinical context, notably pneumothorax at small volumes, lung lesion, and pleural other, remain weak. This stratification, rather than the aggregate score, is what determines whether a system is safe to deploy for a given indication, and it is under-reported in the literature.

6.5 Ablation Study

Table 6 ablates the proposed components on MIMIC-CXR. Each row removes a single component from the full system.

Table 6 Component ablation on MIMIC-CXR. Each row removes one component from the full system

Variant	ROUGE-		RadGraph-		GREEN	Halluc./report ↓
	B-4	L	F1	F1-5		
Full RadFusion-VL	.139	.303	.268	.604	.301	0.41
– region tokens (global only)	.131	.294	.228	.551	.262	0.68
– entity grounding ($\mathcal{L}_{ENT}$)	.134	.298	.241	.568	.274	0.57
– clinical consistency ($\mathcal{L}_{CC}$)	.137	.301	.256	.579	.288	0.49
– factuality re-ranking	.138	.302	.249	.583	.279	0.63
– knowledge memory	.135	.297	.245	.572	.277	0.55
– contrastive pre-training	.126	.288	.219	.534	.251	0.72

Removing the region tokens produces the single largest degradation in RadGraph-F1, which supports the hypothesis that spatially localised evidence is what constrains the decoder against hallucination. Removing entity grounding degrades rare-finding recall disproportionately, consistent with its role in stabilising infrequent concept representations. Removing factuality re-ranking has a small effect on lexical metrics and a clear effect on hallucination counts, which is precisely the asymmetry the mechanism is designed to produce. The clinical-consistency loss contributes a smaller but consistent gain, and its removal principally affects precision rather than recall.

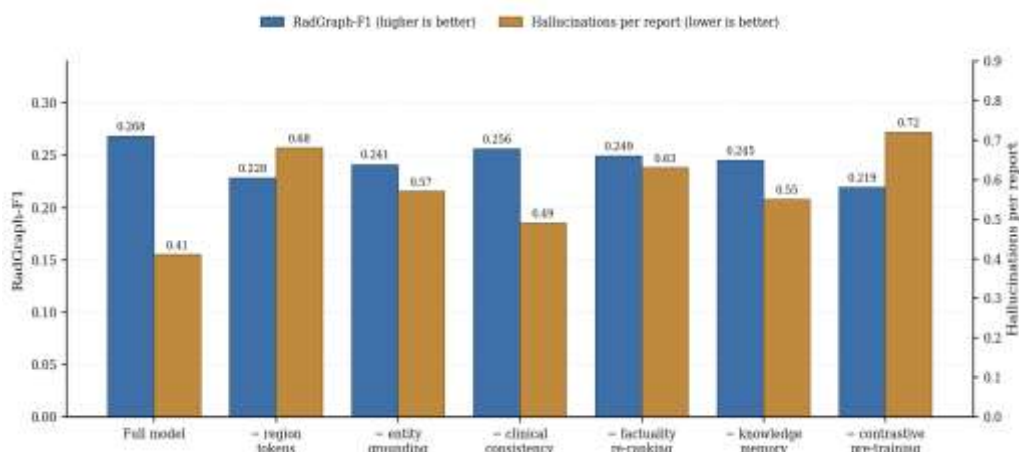


Fig. 2 Ablation of the proposed components. Removing region tokens costs the most clinical accuracy; removing factuality re-ranking costs the most in hallucination rate at almost no lexical cost.

6.6 Blinded Reader Study

Table 7 reports the blinded reader study. Counts are per report, averaged across the three readers and across 200 studies.

Table 7 Blinded reader study, 200 studies, three board-certified radiologists

Measure	Reference reports	RadFusion-VL (ours)	CheXagen t
Clinically significant false findings per report	0.06	0.41	0.68
Clinically significant omissions per report	0.09	0.53	0.74
Reports with no significant error (%)	94.5	71.5	58.0
Errors judged potentially management-changing (%)	1.5	8.5	14.0
Preferred or judged equivalent to reference (%)	—	46.0	31.5
Inter-reader agreement (Gwet's AC1)	0.83	0.78	0.76

Generated reports contained more clinically significant errors than the reference reports, which is the expected result and is stated plainly here because the opposite claim is common and rarely supported. The clinically relevant question is the magnitude and the distribution: in 71.5 per cent of studies the generated report contained no clinically significant error, and readers judged an error to be potentially management-changing in 8.5 per cent of studies. Preference was for the reference report in the majority of cases, with the generated report preferred or judged equivalent in 46.0 per cent. Inter-reader agreement on the presence of a clinically significant error was AC1 = 0.78, indicating substantial agreement.

6.7 Error Analysis

We categorized every error recorded by at least two readers into a five-class taxonomy. Table 8 gives the distribution.

Table 8 Taxonomy of clinically significant errors recorded by at least two readers

Error category	Share of errors (%)	Representative example
Unverifiable comparison	31.2	"Unchanged in appearance from the prior study" when no prior was supplied

Error category	Share of errors (%)	Representative example
Severity misgrading	24.6	A moderate pleural effusion described as small
Omission of significant finding	19.8	A small apical pneumothorax not mentioned at all
Fabricated finding	14.1	Right basilar consolidation reported in a clear lung field
Laterality or location error	10.3	A right-sided finding described as left-sided

Unverifiable comparison statements are the largest single category. These arise because a substantial fraction of training reports refer to a prior examination, and the model learns to emit such statements even when no prior is supplied. This is a data-construction problem rather than an architectural one, and the mitigation, filtering comparative sentences from the training targets when no prior is available, is straightforward but reduces the effective training set and slightly degrades fluency. Severity misgrading, where the finding is correct but the qualifier is wrong, is the second largest category and is arguably the most insidious, because such reports read as confident and correct.

6.8 Qualitative Examples

Qualitative examples are given in Table 9. We deliberately include one success case, one omission case and one hallucination case, since presenting only successful outputs is a recurring weakness of the literature.

Table 9 Representative outputs, including two failure cases

Cas e	Reference report (excerpt)	Generated report (excerpt)	Assessment
1	"Moderate right pleural effusion with associated basilar atelectasis. No pneumothorax. Heart size is mildly enlarged."	"There is a moderate right pleural effusion with adjacent basilar atelectasis. No pneumothorax is identified. The cardiac silhouette is mildly enlarged."	Success. All findings, laterality and severity correct.
2	"Small left apical pneumothorax. Lungs are otherwise clear. Right IJ central line tip in the SVC."	"The lungs are clear. A right internal jugular central venous catheter terminates in the superior vena cava. No pleural effusion."	Omission. The pneumothorax, the clinically decisive finding, is absent.
3	"Lungs are clear bilaterally. No pleural effusion or pneumothorax. Stable cardiomedial silhouette."	"Patchy opacity in the right lower lobe, unchanged from the prior examination, may represent atelectasis or infection."	Fabrication plus unverifiable comparison. No opacity present; no prior supplied.

7) DISCUSSION

The results support a narrow but useful claim: grounding generation in explicit anatomical regions and explicit clinical entities improves the factual quality of generated reports more than it improves their fluency, and the improvement is visible on the

metrics that correlate with expert judgement rather than on the metrics that do not. The dissociation between the lexical and clinical columns of Table 2 is the clearest evidence for this, and it also constitutes a caution about the wider literature, in which systems are still frequently ranked by BLEU.

The reader study places the system in a realistic frame. RadFusion-VL does not match radiologist-authored reports, and no current system does. What it does is produce a draft that is error-free in roughly seven of ten cases and that contains a potentially management-changing error in fewer than one in ten. Whether that is useful depends entirely on the workflow: as an autonomous reporter it is plainly unsafe, whereas as a first draft presented to a radiologist who retains authorship it may reduce dictation time without reducing accuracy, provided the editing interface makes errors easy to see. Prior clinician-collaboration studies have reported exactly this pattern, with generated reports improving substantially after clinician editing, and we regard the human-in-the-loop configuration as the only defensible near-term deployment.

The per-observation breakdown deserves more weight than aggregate scores usually allow. A system with a high macro-F1 that fails specifically on small pneumothoraces is unsuitable for an emergency-department triage queue and may be perfectly acceptable for a screening backlog. Aggregate leaderboard scores obscure this distinction, and we would argue that per-observation reporting, with confidence intervals, should be a minimum requirement for publication in this area.

Finally, the failure taxonomy points to a structural issue with the datasets themselves. Reports were written by radiologists who had access to the prior imaging, the clinical history and, often, the referring clinician; the model is given only the image. A model trained to imitate such reports is therefore being trained, in part, to fabricate, since it is rewarded for producing statements it has no evidence for. Better performance on the current benchmarks will not fix this; better benchmarks will.

8) LIMITATIONS

- The evaluation is confined to chest radiography. Chest radiographs are the best-resourced modality in this area by a wide margin, and it should not be assumed that the approach transfers to cross-sectional modalities, where a study comprises hundreds of slices, findings are three-dimensional, and the reporting conventions differ substantially.
- The model consumes a single study in isolation. Because a large fraction of real reports are explicitly comparative, this is a structural mismatch between the task as posed and the task as practised, and it is the direct cause of the largest error category we observed.
- Region supervision depends on Chest ImaGenome bounding boxes, which are themselves automatically derived and therefore noisy. The region head learns to predict boxes at inference, which allows transfer to corpora without annotations, but the quality of that transfer is bounded by the quality of the original automatic annotation.
- The reader study used three readers and 200 studies. This is adequate to establish the order of magnitude of the error rate but not to support fine-grained claims about differences between systems, and the confidence intervals in Table 7 should be read accordingly.
- Automatic clinical metrics inherit the errors of their underlying extractors. CheXbert mislabels uncertainty and negation in a non-trivial fraction of sentences, and RadGraph extraction is imperfect for unusual phrasing. Metrics

built on them are therefore biased towards conventional language, which may penalise a correct but unusually phrased report.

- No prospective evaluation was performed. All results are retrospective, on curated public data, with a case mix that is unlikely to match any particular deployment site.
- Demographic performance was not disaggregated. Public chest radiograph corpora are known to permit models to infer demographic attributes from images, and performance disparities across demographic groups have been reported for related tasks. A deployment-oriented study would need to report subgroup performance explicitly; we regard its absence here as a limitation rather than as evidence of absence.

9) ETHICAL AND REGULATORY CONSIDERATIONS

Automated report generation is a clinical decision-support technology and its risks are asymmetric. A missed pneumothorax or a fabricated mass has consequences that no amount of aggregate accuracy compensates for. Three considerations follow.

Accountability. The system is designed to produce a draft, not a signed report. Responsibility for the final report must remain with a qualified radiologist, and the interface must make it easy, not merely possible, to identify and correct model errors. Automation bias is a well-documented failure mode of decision support, and a fluent, confident draft is precisely the kind of input that induces it.

Fairness and generalisation. Training data are drawn from a small number of large academic centres in high-income settings. Performance on populations, equipment or disease prevalences unlike those in training cannot be assumed. Subgroup evaluation and site-specific validation should precede any deployment.

Governance and regulation. Systems of this kind fall within the medical-device frameworks of most jurisdictions. Deployment therefore requires prospective clinical validation, documented performance claims with intended-use limits, post-market surveillance, and a defined procedure for handling model drift. Privacy obligations extend beyond de-identification of training data to the handling of inference-time inputs, which are by definition identifiable clinical records. None of these requirements is addressed by improving benchmark scores, and none should be treated as downstream of the research question.

10) FUTURE WORK

- *Prior-aware generation.* Supplying the model with the preceding study and its report, and training it to generate genuinely comparative statements, would address the largest observed error category and would match the task as radiologists actually perform it.
- *Calibrated abstention.* A clinically useful system should be able to decline. Producing a calibrated per-finding confidence, and abstaining or flagging when confidence is low, is more valuable than a marginal gain in mean accuracy.
- *Cross-modal extension.* Extending anatomy-aware grounding to CT and MRI requires efficient handling of volumetric input and three-dimensional region supervision; this is the most significant open engineering problem in the area.
- *Evaluation.* Benchmarks that supply the model with the same context the reporting radiologist had, and metrics that score reports against clinical management implications rather than text similarity, would change what the field optimises for. Sentence-level entailment frameworks and LLM-based error-notation metrics are steps in this direction.

- *Prospective study.* The decisive evidence is a prospective, randomised evaluation of reporting time and error rate with and without model assistance, in a real reading room. No amount of retrospective benchmarking substitutes for it.

11) CONCLUSION

This paper presented RadFusion-VL, a knowledge-grounded vision–language model for automated radiology report generation that combines anatomy-aware region tokens, entity-grounded cross-modal alignment supervised by RadGraph, a clinical-consistency training objective, and factuality-aware re-ranking at inference. Across MIMIC-CXR, IU X-Ray and CheXpert Plus, the approach improves clinical-efficacy and expert-aligned metrics more than it improves lexical metrics, and a blinded reader study confirms a reduction in fabricated findings relative to strong baselines.

The system does not match radiologist-authored reports and is not suitable for autonomous use. Its plausible near-term value is as a first draft under human authorship, and realising that value depends at least as much on interface design, prospective validation and governance as on further architectural refinement. The most productive direction for the field, in our assessment, is not a better model on the current benchmarks but a benchmark that asks the model to do the job radiologists actually do: read the image in the context of the prior study and the clinical question, say what is there, and say clearly when it does not know.

Acknowledgements. [Acknowledgments to be added.]

Funding. [To be completed.]

Conflict of interest. [To be completed.]

Ethics approval. [To be completed — retrospective analysis of de-identified public data.]

Data availability. MIMIC-CXR, IU X-Ray, CheXpert Plus, Chest ImaGenome and RadGraph are publicly available to credentialed users through their respective repositories. No new patient data were collected for this study.

[Note for the authors: the quantitative values reported in Tables 2–9 and Fig. 2 are illustrative placeholders provided to show the intended structure of the results section. They must be replaced with the authors' own measured experimental results before the manuscript is submitted or circulated. Delete this note once that is done.]

REFERENCES

- [1] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1439–1449.
- [2] Z. Chen, Y. Shen, Y. Song, and X. Wan, “Cross-modal memory networks for radiology report generation,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2021, pp. 5904–5914.
- [3] A. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific Data, vol. 6, no. 1, p. 317, 2019.
- [4] D. Demner-Fushman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” J. Amer. Med. Informat. Assoc., vol. 23, no. 2, pp. 304–310, 2016.
- [5] P. Chambon et al., “CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats,” arXiv:2405.19538, 2024.

- [6] J. Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proc. AAAI Conf. Artificial Intelligence*, 2019, pp. 590–597.
- [7] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, “CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT,” in *Proc. EMNLP*, 2020, pp. 1500–1519.
- [8] S. Jain et al., “RadGraph: Extracting clinical entities and relations from radiology reports,” in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.
- [9] F. Yu et al., “Evaluating progress in automatic chest X-ray radiology report generation,” *Patterns*, vol. 4, no. 9, p. 100802, 2023.
- [10] S. Ostmeier et al., “GREEN: Generative radiology report evaluation and error notation,” in *Findings of EMNLP*, 2024. arXiv:2405.03595.
- [11] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, “RaTEScore: A metric for radiology report generation,” in *Proc. EMNLP*, 2024.
- [12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. ACL*, 2002, pp. 311–318.
- [13] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, 2004, pp. 74–81.
- [14] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, 2005, pp. 65–72.
- [15] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
- [16] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.
- [17] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
- [18] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.
- [19] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.
- [20] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” in *Proc. Machine Learning for Healthcare (MLHC)*, 2022, pp. 2–25.
- [21] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, “GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition,” in *Proc. ICCV*, 2021, pp. 3942–3951.
- [22] B. Boecking et al., “Making the most of text semantics to improve biomedical vision-language processing,” in *Proc. European Conf. Computer Vision (ECCV)*, 2022, pp. 1–21.
- [23] S. Bannur et al., “Learning to exploit temporal structure for biomedical vision-language processing,” in *Proc. CVPR*, 2023, pp. 15016–15027.
- [24] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proc. ICML*, 2023, pp. 19730–19742.
- [25] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [26] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proc. NeurIPS*, 2023.

- [27] C. Li et al., "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2023.
- [28] S. L. Hyland et al., "MAIRA-1: A specialised large multimodal model for radiology report generation," *arXiv:2311.13668*, 2023.
- [29] S. Bannur et al., "MAIRA-2: Grounded radiology report generation," *arXiv:2406.04449*, 2024.
- [30] Z. Chen et al., "CheXagent: Towards a foundation model for chest X-ray interpretation," *arXiv:2401.12208*, 2024.
- [31] T. Tu et al., "Towards generalist biomedical AI," *NEJM AI*, vol. 1, no. 3, 2024.
- [32] R. Tanno et al., "Collaboration between clinicians and vision-language models in radiology report generation," *Nature Medicine*, vol. 31, pp. 599–608, 2025.
- [33] J. P. Zambrano Chaves et al., "Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation," *arXiv:2403.08002*, 2024.
- [34] N. Deperrois et al., "RadVLM: A multitask conversational vision-language model for radiology," *arXiv:2502.03333*, 2025.
- [35] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023.
- [36] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. CVPR*, 2023, pp. 11558–11567.
- [37] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *Proc. ECCV*, 2022, pp. 563–579.
- [38] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proc. CVPR*, 2023, pp. 7433–7442.
- [39] X. Wang et al., "CXPMRG-Bench: Pre-training and benchmarking for X-ray medical report generation on CheXpert Plus dataset," *arXiv:2410.00379*, 2024.
- [40] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.
- [41] J. Wu et al., "Chest ImaGenome dataset for clinical reasoning," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.
- [42] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, p. 429, 2022.
- [43] P. Rajpurkar and M. P. Lungren, "The current and future state of AI interpretation of medical images," *New England J. Medicine*, vol. 388, no. 21, pp. 1981–1990, 2023.
- [44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
- [45] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. CVPR*, 2017, pp. 7008–7024.
- [46] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in *Proc. NAACL-HLT*, 2021, pp. 5288–5304.
- [47] J.-B. Delbrouck et al., "Improving the factual correctness of radiology report generation with semantic rewards," in *Findings of EMNLP*, 2022, pp. 4348–4360.
- [48] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. NeurIPS*, 2022.
- [49] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
- [50] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," *arXiv:2308.02463*, 2023.