Review

Current Landscape of Automatic Radiology Report Generation with Deep Learning: An Exploratory Systematic Review

Patricio Melendez-Rojas *, Jaime Jamett-Rojas, María Fernanda Villalobos-Dellafiori, Pablo R. Moya, Alejandro Veloz-Baeza

Posted Date: 3 November 2025

doi: 10.20944/preprints202511.0010.v1

Keywords: deep learning; digital images; natural language processing; radiology; transformers

 doi:10.20944/preprints202511.0010.v1

Review

Current Landscape of Automatic Radiology Report Generation with Deep Learning: An Exploratory Systematic Review

Patricio Meléndez-Rojas [1,2,*], Jaime Jamett-Rojas [1], M. Fernanda Villalobos-Dellafiori [2], Pablo R. Moya [3,4] and Alejandro Veloz-Baeza [5]

[1] Health Sciences and Engineering, Universidad de Valparaíso, Valparaíso, Chile
[2] Faculty of Dentistry, Universidad Andres Bello, Viña del Mar, Chile
[3] Centro Interdisciplinario de Neurociencias de Valparaíso (CINV), Universidad de Valparaíso, Valparaíso, Chile
[4] Institute of Physiology, Faculty of Sciences, Universidad de Valparaíso, Valparaíso, Chile
[5] Faculty of Engineering, Universidad de Valparaiso, Valparaíso, Chile
* Correspondence: patricio.melendez@postgrado.uv.cl

Abstract

Automatic radiology report generation (ARRG) has emerged as a promising application of deep learning (DL) with the potential to alleviate reporting workload and improve diagnostic consistency. However, despite rapid methodological advances, the field remains technically fragmented and not yet mature for routine clinical adoption. This systematic review maps the current ARRG research landscape by examining DL architectures, multimodal integration strategies, and evaluation practices from 2015 to April 2025. A PRISMA-compliant search identified 89 eligible studies, revealing a marked predominance of chest radiography datasets (87.6%), largely driven by their public availability and the accelerated development of automated tools during the COVID-19 pandemic. Most models employed hybrid architectures (73%), particularly CNN–Transformer pairings, reflecting a shift toward systems capable of combining local feature extraction with global contextual reasoning. Although these approaches have achieved measurable gains in textual and semantic coherence, several challenges persist, including limited anatomical diversity, weak alignment with radiological reasoning, and evaluation metrics that insufficiently reflect diagnostic adequacy or clinical impact. Overall, the findings indicate a rapidly evolving but clinically immature field, underscoring the need for validation frameworks that more closely reflect radiological practice and support future deployment in real-world settings.

Keywords: deep learning; digital images; natural language processing; radiology; transformers

1. Introduction

The interpretation of medical images and the generation of radiological reports constitute a core component of diagnostic assessment, treatment planning, and ongoing patient monitoring [1,2]. Producing a coherent and clinically meaningful report requires not only the accurate recognition of imaging findings but also their integration into a structured diagnostic narrative, a process that demands years of specialized training and contributes to workload pressures in radiology departments [1–3]. These constraints, together with the risk of inter-observer variability, have motivated growing interest in computational systems capable of supporting or partially automating the reporting process [1–3].

Deep learning (DL) has become the predominant paradigm in automatic radiology report generation (ARRG) [1], commonly implemented through encoder–decoder architectures in which convolutional neural networks (CNNs) extract visual representations and text-based decoders

 doi:10.20944/preprints202511.0010.v1

generate the final report [3]. Early works predominantly relied on recurrent neural networks (RNNs), including long short-term memory (LSTM) models [1], whereas recent advances have introduced attention-based mechanisms such as Transformers [4] and multimodal contrastive frameworks such as contrastive language-image pretraining (CLIP) [5]. Further developments include domain knowledge-guided strategies [1,6], attention-based architectures [7], reinforcement learning [2], large language models (LLMs) [8,9], and hybrid approaches that integrate multiple mechanisms to improve clinical accuracy [1].

However, despite these methodological advances, the current body of evidence remains fragmented, with substantial gaps in clinically aligned validation, semantic faithfulness, and the integration of structured medical knowledge into report generation pipelines [1,2,5]. In this context, most published ARRG models demonstrate promising linguistic performance but remain in a stage of limited clinical readiness, showing insufficient validation for deployment in routine radiological workflows.

The advancement of ARRG relies on specialized datasets comprising medical images paired with textual reports [1,5], with public benchmarks such as MIMIC-CXR and IU-Xray frequently supporting this line of research [1,7]. Yet, evaluating model performance remains challenging due to the heterogeneity of metrics: some focus on textual similarity (e.g., BLEU, ROUGE, BERTScore), while others estimate the correctness of clinical findings (e.g., AUC, F1 score), often lacking standardized clinical validation [1,5–7]. Although prior reviews have addressed selected aspects of ARRG or specific architectures such as Transformers and multimodal methods [3–5,11], the rapid proliferation of DL-based systems has produced a fragmented landscape that complicates the global understanding of methodological progress and its alignment with clinical readiness.

Accordingly, the main objective of this systematic review is to explore the current research landscape on ARRG using DL, describing their key characteristics, methodological approaches, data sources, evaluation strategies, and principal findings, while also examining their alignment with the criteria required for future real-world adoption.

2. Materials and Methods

A comprehensive PRISMA-compliant search was conducted in the following electronic databases: IEEE Xplore, ACM Digital Library, PubMed/MEDLINE, Scopus, and Web of Science (WoS) Core Collection. This review included original English-language research articles from peer-reviewed journals published between 2015 and April 2025. Eligible studies were required to address the automatic generation of radiology reports using deep learning (DL) architectures, including but not limited to CNNs, RNNs, Transformers, graph neural networks (GNNs), or hybrid models. Studies additionally had to employ multimodal input data, typically pairing imaging data with a text-generation component, with or without additional structured information, as a core element of the report generation pipeline. Exclusion criteria comprised studies outside the radiology domain, non-original publications (such as reviews or editorials), and works lacking sufficient methodological or outcome detail for full assessment. Grey literature was not considered in this review.

The detailed search strategies applied to each database are presented in Table 1. After removing duplicates, titles and abstracts were independently screened by the first author (MP) and the second author (JJ) according to the predefined eligibility criteria. Full-text versions of potentially relevant publications were retrieved and independently evaluated by MP and JJ for final inclusion. Any disagreements arising at either screening stage were resolved by the third reviewer (VM).

Data extraction focused on key study characteristics, including study objectives, radiological domain, datasets used, input modalities, DL architectures and methodological details, report generation pipeline characteristics, and evaluation metrics. Bibliometric information (authors, publication year, country, and publication source), reported limitations, and suggested future research directions were also recorded. The extraction form was jointly developed by MP and JJ, who independently completed and iteratively refined the dataset until consensus was reached. Studies

 doi:10.20944/preprints202511.0010.v1

were subsequently grouped according to the DL methodology employed, and their main characteristics and findings were synthesized narratively.

Table 1. Search strategies adapted by database.

Database	Field	Search Expression	Results
PubMed	[tiab]	("Convolutional Neural Network*" OR CNN OR "Recurrent Neural Network*" OR RNN OR LSTM OR GRU OR Transformer OR Transformers OR "Attention Mechanism" OR "Encoder Decoder" OR "Sequence to Sequence" OR "Graph Neural Network*" OR GNN OR GCN OR GAT OR "Deep Learning" OR "Neural Network" OR "Neural Networks") AND (Radiology OR Radiolog* OR "Medical Imag*" OR "Diagnostic Imag*" OR X-ray OR CT OR MRI OR PET) AND ("Report Generation" OR "Text Generation" OR "Narrative Generation" OR "Automatic Report*" OR "Clinical Report*" OR "Medical Report*")	158
Scopus	-	Same expression as PubMed, without specific field restriction	259
Web of Science	TS=	Same Boolean expression adapted to the TS= field for topic-based search	217
IEEE Xplore	-	Same Boolean expression adjusted to the syntax requirements of the respective database	79
ACM DL	-	Same Boolean expression adjusted to the syntax requirements of the respective database	301

In PubMed, the [tiab] field was used to restrict the search to title and abstract. In WoS, the TS= field was used for topic-based search. Search expressions were syntactically adapted to each database.

Risk of bias arising from missing results was addressed qualitatively. Selective outcome reporting was assessed during the extraction process, and completeness of reporting was evaluated using the TRIPOD-LLM guideline [12]. Due to the exploratory nature of this review, no formal quantitative assessment of publication bias was conducted; however, potential reporting-related sources of bias were considered in the synthesis. The review protocol was registered in PROSPERO under the registration number CRD420251044453.

3. Results

A PRISMA flow chart of the screening and selection process is presented in Figure 1, and Table 1 summarizes the number of articles retrieved from each database. A total of 89 studies met the inclusion criteria. Research activity in this field has intensified markedly in recent years: one eligible study was published in 2020, followed by five in 2021, nineteen in 2022, fourteen in 2023 and twenty-nine in 2024, with a slight decline to twenty-one in the partial 2025 dataset (Figure 2). Assessment using the TRIPOD-LLM checklist showed that while most studies reported performance metrics in detail, essential elements such as participant description, sample size justification, and external validation were frequently absent, and none of the reviewed articles achieved full compliance.

 doi:10.20944/preprints202511.0010.v1

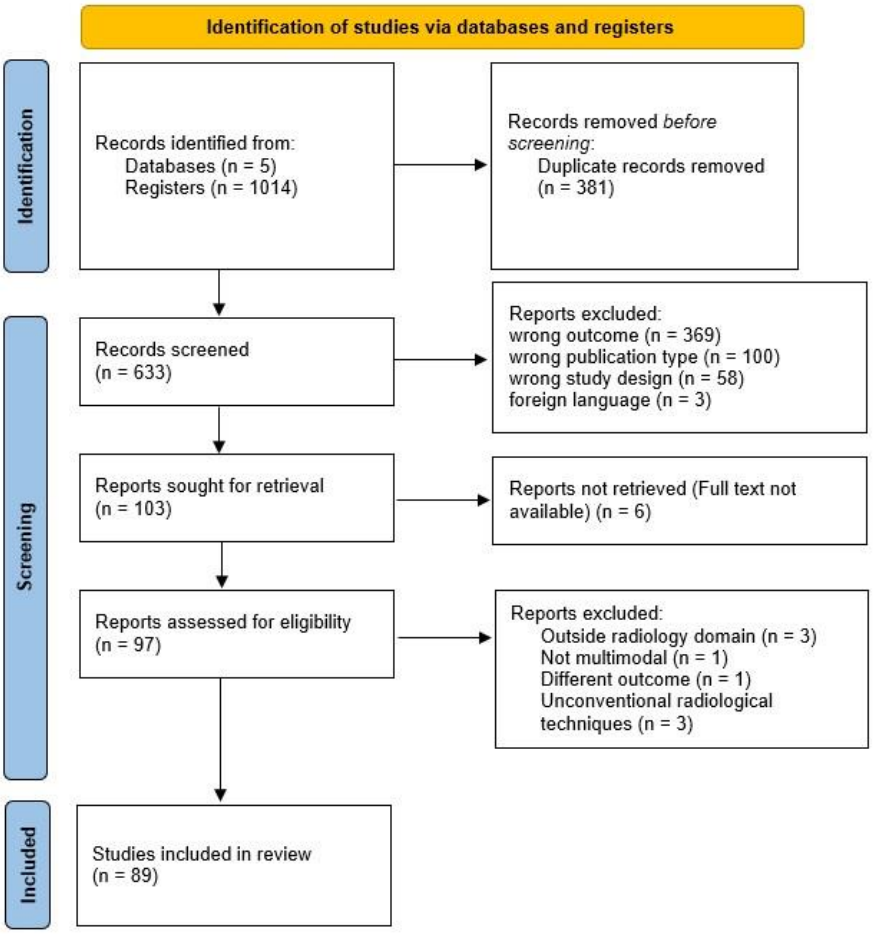

Figure 1. PRISMA diagram [13] for the systematic review demonstrating the search results, included and excluded studies.

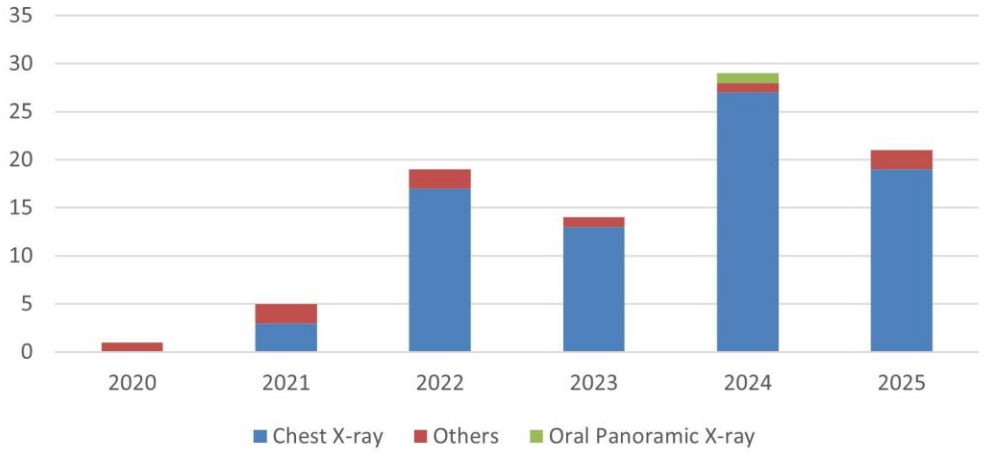

Figure 2. Number of publications by field of clinical focus and year.

Regarding 78. out of 89 studies (87.6%) using chest X-ray (CXR) datasets, primarily from publicly available repositories. Other anatomical regions such as brain CT/MRI, abdominal CT, spinal imaging, and oral/maxillofacial radiology were represented only sporadically, with four or fewer publications each (Figure 2). This distribution highlights that although ARRG has advanced rapidly, its development remains largely restricted to thoracic imaging.

 doi:10.20944/preprints202511.0010.v1

Geographic distribution further reveals that ARRG research is concentrated almost exclusively in a small number of countries. As shown in Figure 3, China accounts for the majority of publications (n=56), followed at a distance by India (n=9), Pakistan (n=4), and Brazil (n=3), while all remaining contributing nations report no more than two studies each. This pattern reflects structural disparities in AI research capacity and emphasizes that most methodological innovation in ARRG is being driven by institutions with access to large-scale datasets and high computational resources, predominantly located in high-resource settings.

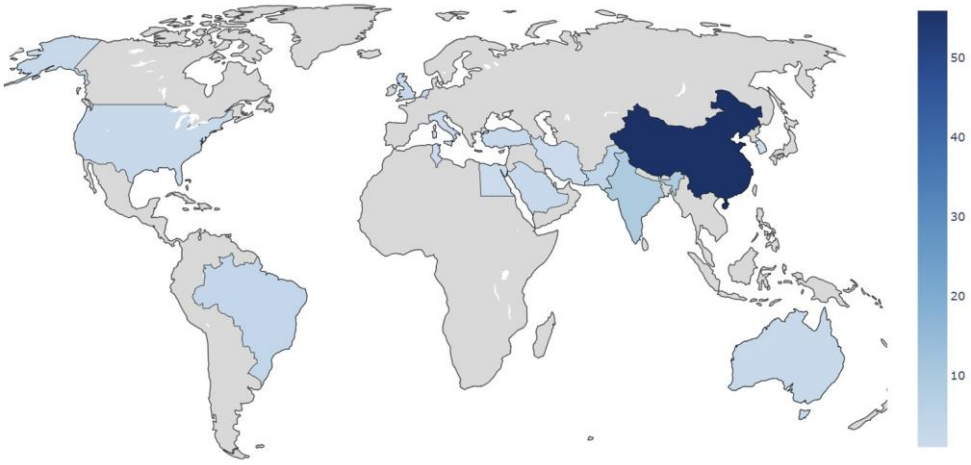

Figure 3. Geographic distribution of the included studies by country of origin. China contributes the largest share of publications (n=56), followed by India (n=9), Pakistan (n=4), and Brazil (n=3).

From a methodological perspective, hybrid DL architectures were the most frequently represented approach. Studies most commonly combined convolutional encoders with Transformer-based language modeling, followed by pure Transformer-based strategies, more complex multi-hybrid approaches, and finally CNN–RNN pipelines, which continue to appear but at a reduced rate. Together, these trends indicate a clear shift toward architectures that integrate localized feature extraction with broader contextual reasoning. This architectural distribution reflects not only technical preferences but also the gradual movement toward models designed to encode both localized saliency and global diagnostic context. To contextualize how frequently each family of models appears in the literature, Figure 4 summarizes the relative prevalence of the main deep learning approaches represented in the included studies.

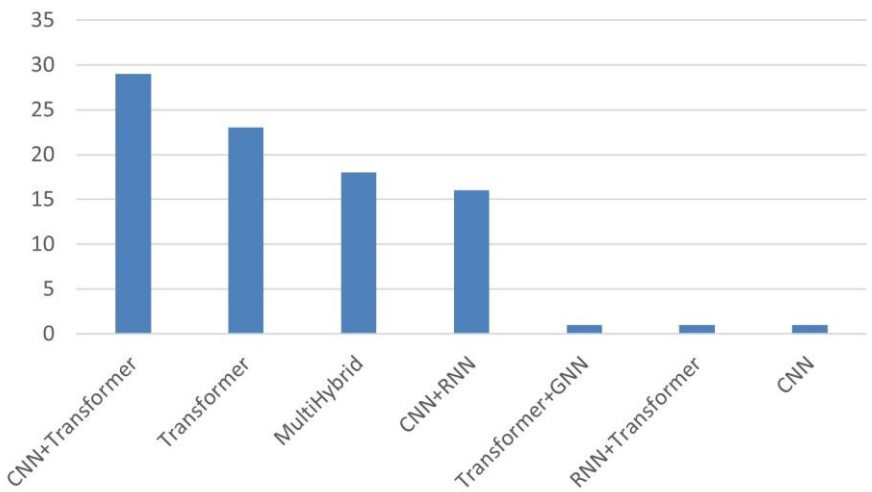

Figure 4. Number of publications by deep learning method.

 doi:10.20944/preprints202511.0010.v1

To facilitate comparison across methodological strategies, the studies were categorized into four architectural groups according to their prevalence in the included literature: (i) CNN–Transformer hybrid architectures, (ii) purely Transformer-based methods, (iii) multi-hybrid combinations integrating multiple DL paradigms, and (iv) traditional CNN–RNN encoder–decoder pipelines. In addition to architectural prevalence, performance across the included studies was assessed using BLEU-1 as the most frequently reported textual similarity metric. As illustrated in Figure 5, BLEU-1 values show substantial variability across models. These categories structure the subsequent analysis of performance characteristics and evaluation approaches (Figures 4 and 5).

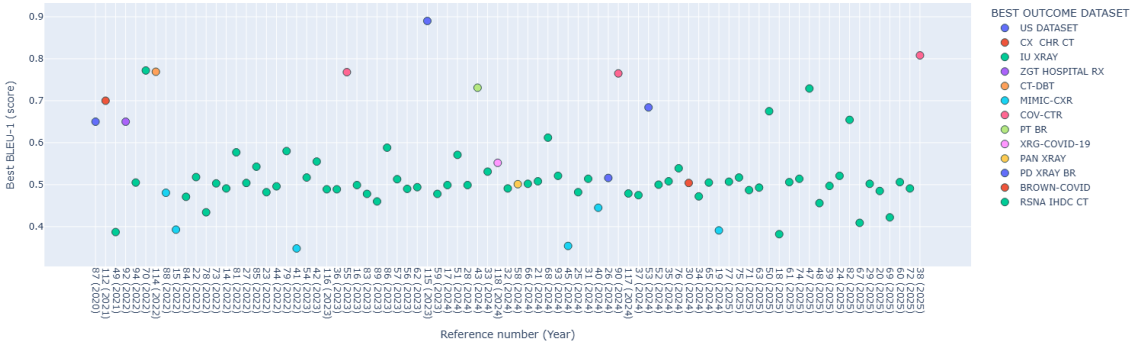

Figure 5. Distribution of BLEU-1 scores from the included studies. Each point on the vertical axis represents the BLEU-1 value obtained by a study; studies that did not report this metric are omitted. Notably, the highest scores were achieved on datasets other than the common IU-Xray and MIMIC-CXR benchmarks.

3.1. CNN + Transformers

ARRG methods have increasingly adopted encoder-decoder architectures, leveraging CNNs for visual encoding and Transformers networks for language decoding [14–16]. This represents a shift from earlier approaches utilizing RNNs like LSTMs [14,16,17], with Transformers offering superior capacity to model long-range dependencies and process information in parallel, resulting in richer contextual representations [15,18]. Widely cited examples of this framework include the Memory-driven Transformer (R2Gen) [14,18–26] and its variant, the Cross-modal memory network (R2GenCMN) [16,18,19,24–28], which enhance information flow and cross-modal alignment. Other approaches such as RATCHET ncorporate a standard Transformer decoder guided by CNN-derived features [14,29], while additional enhancements include relation memory units and cross-modal memory matrices [30].

A persistent challenge for these models is the inherent data imbalance in medical datasets, where normal findings vastly outnumber abnormal ones [14,20–22,30]. To mitigate this, contrastive learning strategies have been introduced, as in the Contrastive Triplet Network (CTN), which improves the representation of rare abnormalities by contrasting visual and semantic embeddings [14]. Integration of medical prior knowledge is another common enhancement, through knowledge graphs [14,19,21], disease tags [19,23,31], or anatomical priors [23], helping to guide generation toward clinically meaningful structures [32]. Additional refinements include organ-aware decoders [31], multi-scale feature fusion [17,33], and adaptive mechanisms that dynamically modulate the contribution of visual and semantic inputs [22]. Ablation studies consistently confirm the value of these specialized components in improving performance [16,33–35].

Evaluations on public datasets, notably IU X-ray [14,15,23,25,28,33,36] and MIMIC-CXR [14,15,23,25,28,33,36], demonstrate that CNN-Transformer based methods achieve state-of-the-art (SOTA) results across conventional natural language generation metrics including BLEU, METEOR, ROUGE-L, and CIDEr [14,16,21,28,30,31,33,36]. Reported improvements include better abnormality description, higher sentence coherence, and enhanced medical correctness [14,15,23,24,34,37], with successful applications extending to cranial trauma and polydactyly reporting [18,26]. Nevertheless,

 doi:10.20944/preprints202511.0010.v1

these models still encounter difficulty in representing uncertainty, severity, and extremely rare pathologies [34], indicating that although their textual fidelity has advanced considerably, their clinical interpretability and readiness for deployment remain limited.

3.2. Transformers

Transformer-based architectures have become a potent solution for ARRG, offering significant advancements over traditional CNN–RNN and LSTM approaches [38–43]. Their primary advantage lies in the ability to model long-range dependencies, which is critical for radiological reporting, where multiple anatomical and pathological findings must be described coherently within a single narrative [38,40,43,44]. These systems are typically structured as encoder–decoder framework [38–40,45–47] or as pure Transformer-based designs [44]. Many implementations leverage pretrained models such as Vision Transformers (ViT) or Swin Transformer for image encoding, paired with language models like GPT-2 or BERT for generation, improving performance in scenarios constrained by limited medical data [38,42,43,45,46,48–52].

Attention mechanisms play a central role in integrating multimodal features [39,40,42,43,53,54]. Cross-attention modules allow fine-grained alignment between visual and textual embeddings, improving structural coherence and semantic grounding [38,39,45,46,51]. Additional optimization strategies include graph-based fusion [38] multi-feature enhancement modules [54], and specialized mechanisms to prioritize rare or diagnostically relevant content [48]. Memory augmentation [38–40,42,51,55,56], knowledge integration through factual priors or graphs [39,51,54–57] and object-level feature extraction [58] further enhance semantic accuracy. Term-weighting and vocabulary-masking strategies reduce overreliance on frequent normal descriptors, improving recall of abnormal findings [44].

Transformers are typically evaluated using standard natural language generation metrics such as BLEU, ROUGE, and METEOR [38,42,44–46,51,52,54,56,59,60], as well as semantic similarity measures including BERTScore and CheXbert [46,49,51,58,60]. Some studies incorporate clinical validation via tools such as RadGraph or expert assessment [41,46,47,60], reflecting a growing emphasis on clinically meaningful correctness. However, despite these advances, most models still rely on surrogate textual similarity metrics and have not yet demonstrated consistent alignment with radiological reasoning in real-world settings, indicating that their clinical maturity remains preliminary.

3.3. Multihybrid

There is a clear trend toward multi-hybrid architectures that combine different DL paradigms and integrate external medical knowledge to enhance semantic alignment and clinical relevance [61,62]. Most approaches continue to employ an encoder–decoder structure in which CNNs such as ResNet [61–66] or VGG19 [61,67,68]) function as visual encoders, while LSTM-based [61–63,65,66,69] or hierarchical RNN decoders [68,70] generate the textual output, with Transformers increasingly incorporated to improve long-term dependency modeling [61,62,64–73].

A central design objective of these hybrid configurations is to improve cross-modal alignment between image regions and textual concepts. To this end, memory networks are widely used to store or retrieve image–text correspondences [61,62,64,66,69,74,75], align medical terminology with localized visual features [64,72,75], or stabilize representations during generation [61,62,64,66,67,72,73,75,76]. Contrastive learning is also frequently adopted to reinforce semantic distinction between positive and negative image–text pairs, improving the detection of clinically relevant abnormalities [63,66,67,75]. External knowledge sources, including knowledge graphs [66,67,74,75], curated medical vocabularies [72,75], and retrieval-augmented mechanisms drawing from similar past reports [62,64,72,73,75], are increasingly used to embed domain expertise into the generation process.

Some models also emulate elements of radiological reasoning by incorporating multi-expert or multi-stage workflows [65] or by implementing strategies that first localize salient regions and then

 doi:10.20944/preprints202511.0010.v1

generate narratively coherent text [72]. Others address data imbalance and background noise by refining lesion-level representations through denoising or saliency-aware mechanisms [67,72,77], or by prioritizing diagnostically abnormal content through report-level reordering [70]. The introduction of contextual embeddings derived from large language models such as BERT further improves lexical richness and contextual fidelity [68,70,73].

Performance across this category is typically benchmarked using IU-Xray [61–64,66–76] and MIMIC-CXR datasets [61–67,69–76]. Reported gains are consistent across BLEU [61–72,74–77], METEOR [61–67,69,71,72,74,75,77,78], ROUGE-L [61–72,74–77], and CIDEr [64,68,70–72,74–77], with qualitative analyses [69,70,73] and ablation studies [66,71,73] confirming the contribution of hybrid components to improved fluency, abnormality description, and semantic correctness [65,69,71]. However, despite these gains, clinical deployment remains limited, as most evaluation frameworks still prioritize textual similarity rather than diagnostic alignment or interpretive reasoning.

3.4. CNN + RNN Architectures

ARRG frequently employs encoder-decoder architectures [78–84], commonly consisting of a CNN encoder to extract visual features from medical images [74,79–85], and a RNN decoder, often a LSTM or Gated Recurrent Unit (GRU), to generate the report text sequentially [74,78,79,81–84,86–88]. This framework seeks to translate visual representations into descriptive narratives [78,79,84,87]. Performance is commonly evaluated using standard NLG metrics such as BLEU [74,78,80,82–85,88–91], ROUGE (especially ROUGE-L) [78,80,82–85,88,90,91], CIDEr [80,83,85,88–90], and METEOR [80,82,84,88,90] with some studies supplementing text-based evaluation with diagnostic accuracy pipelines or automated clinical assessment tools such as CheXpert [80,82,91].

Several CNN+RNN-based methods have reported competitive or even superior performance across multiple benchmarks [79,81,82,88,89,91]. For instance, a CNN-LSTM model incorporating attention achieved a BLEU-4 of 0.155 on MIMIC-CXR [79], while the G-CNX network, combining ConvNeXtBase with a GRU decoder, obtained BLEU-1 scores of 0.6544 on IU-Xray and 0.5593 on ROCOv2 [82]. Similarly, the HReMRG-MR method, based on LSTMs with reinforcement learning, demonstrated improvements over several baselines on both IU-Xray and MIMIC-CXR [88]. Additional architectures targeting specialized reporting tasks, such as proximal femur fracture assessment in Dutch, have also reported strong performance [89].

However, despite these promising results, CNN+RNN models are limited by their sequential decoding nature, which restricts long-range contextual reasoning and reduces their ability to handle complex radiological narratives. As a result, while these architectures remain relevant in benchmarking and resource-constrained settings, their suitability for clinically aligned reporting is inherently limited relative to Transformer-based or hybrid approaches.

3.5. Others

Three studies employed architectures that did not fit into the previously defined categories due to distinctive design choices targeting specialized aspects of ARRG. The first approach replaces free-text generation with structured output by learning question-specific representations using tailored CNNs and MobileNets, which are subsequently classified with SVMs rather than decoded through a sequence generator [92]. This method demonstrated superior performance on the ImageCLEF2015 Liver CT annotation task, suggesting that task-adapted feature extraction can outperform shared representations in structured reporting problems. A second framework integrates a ViT encoder with a hierarchical LSTM decoder, augmented by a MIX-MLP module for multi-label classification and a POS-SCAN co-attention mechanism to fuse semantic and visual priors [93]. This hybrid configuration leverages label-aware alignment to improve anomaly identification and text coherence, with ablation studies confirming the contribution of its integrated components. The third approach incorporates a knowledge graph derived from disease label co-occurrence statistics, combining DenseNet-based visual features with a Transformer text encoder and a GNN reasoning layer before final report generation [94], thereby enabling structured medical knowledge to inform the decoding process.

 doi:10.20944/preprints202511.0010.v1

Table 2. Summary of key characteristics, methodologies, and reported metrics for the included studies.

Year	Ref.	Radiological Domain	Datasets (DS) Used	Deep Learning Method	Architecture	Best BLEU-1	Best Outcome DS
2020	[87]	Ultrasound (US) (gallbladder, kidney, liver), Chest X-ray	US image dataset (6,563 images). IU-Xray (7,470 images, 3,955 reports)	CNN, RNN	SFNet (Semantic fusion network). ResNet-50. Faster RCNN. Diagnostic report generation module using LSTM	0.65	US DATASET
2021	[112]	Chest CT	COVID-19 CT dataset (368 reports, 1,104 CT images). CX-CHR dataset (45,598 images, 28,299 reports), 12 million external medical textbooks	Transformer, CNN	Medical-VLBERT (with DenseNet-121 as backbone)	0.70	CX CHR CT
2021	[49]	Chest X-ray	IU-Xray	Transformer	2 CDGPT2 (Conditioned Distil Generative Pre-trained Transformer)	0.387	IU-Xray
2021	[113]	Chest X-ray	MIMIC-CXR (377,110 images, 227,835 reports). IU-Xray	CNN, Transformer	BERT-base	*0.126 (BLEU-4)	MIMIC-CXR
2021	[92]	Liver CT	Liver CT annotation dataset from ImageCLEF 2015 (50 patients)	CNN	MobileNet-V2	0.65	ZGT HOSPITAL RX
2021	[80]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, GAN	Encoder with two branches (CNN based on ResNet-152 and MLC). Hierarchical LSTM decoder with multi-level attention and a reward module with two discriminators.	*0.148 (BLEU-4)	MIMIC-CXR
2022	[91]	Proximal Femur Fractures (X-ray)	Main dataset: 4,915 cases with 11,606 images and reports. Language model dataset: 28,329 radiological reports	CNN, RNN	DenseNet-169 for classification. Encoder-Decoder for report generation. GloVe for language modeling	0.65	MAIN DATASET
2022	[94]	Chest X-ray	MIMIC-CXR. IU-Xray	GNN, Transformer	Custom framework using Transformer for generation module	0.505	IU-Xray
2022	[70]	Chest X-ray	IU-Xray	CNN, RNN, Transformer	CNN VGG19 network (feature extraction). BERT (language generation). DistilBERT (perform sentiment)	0.772	IU-Xray
2022	[114]	Liver CT and kidney, DBT (Digital Breast Tomosynthesis)	ImageNet (25,000 images). CT abdomen and mammography images (750 images). CT and DBT medical images (150 images)	RNN, CNN	MLTL-LSTM model (Multi level transfer learning framework with a long short-term-memory model)	0.769	CT-DBT
2022	[88]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN	HReMRG-MR (Hybrid reinforced report generation method with m-linear attention and repetition penalty mechanism)	0.4806	MIMIC-CXR
2022	[15]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	CvT2DistilGPT2	0.4732	IU-Xray
2022	[64]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	DenseNet (encoder). LSTM or Transformer (decoder). ATAG (Attributed abnormality graph) embeddings. GATE (gating mechanism)	**0.323 (BLEU AVG)	IU-Xray
2022	[84]	Chest X-ray	IU-Xray	CNN, RNN	AMLMA (Adaptive multilevel multi-attention)	0.471	IU-Xray
2022	[22]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	MATNet (Multimodal adaptive transformer)	0.518	IU-Xray

 doi:10.20944/preprints202511.0010.v1

2022	[78]	Chest X-ray	IU-Xray	CNN, RNN, Attention Mechanism	RCLN model (combining CNN, LSTM, and multihead attention mechanism). Pre-trained ResNet-152 (image encoder)	0.4341	IU-Xray
2022	[73]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	VTI (Variational topic inference) with LSTM-based and Transformer-based decoders	0.503	IU-Xray
2022	[14]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	CTN built on Transformer architecture	0.491	IU-Xray
2022	[81]	Chest X-ray	IU-Xray	CNN, RNN, Reinforcement Learning	CADxReport (VGG19, HLSTM with co-attention mechanism and reinforcement learning)	0.577	IU-Xray
2022	[27]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	CAMANet (Class activation map guided attention network)	0.504	IU-Xray
2022	[85]	Chest X-ray	IU-Xray	CNN, RNN, Attention Mechanism	CheXPrune (encoder-decoder architecture with VGG19 and hierarchical LSTM)	0.5428	IU-Xray
2022	[23]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer, VAE	Prior Guided Transformer. ResNet101 (visual feature extractor). Vanilla Transformer (baseline)	0.482	IU-Xray
2022	[44]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	Pure Transformer-based Framework (custom architecture)	0.496	IU-Xray
2022	[79]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN	VGG16 (visual geometry group CNN network). LSTM with attention	0.580	IU-Xray
2022	[41]	Chest X-ray	MIMIC-CXR. CheXpert	Transformer	Meshed-memory augmented transformer architecture with visual extractor using ImageNet and CheXpert pre-trained weights	0.348	MIMIC-CXR
2023	[54]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	MFOT (Multi-feature optimization transformer)	0.517	IU-Xray
2023	[42]	Chest X-ray	IU-Xray	Transformer	TrMRG (Transformer Medical Report Generator) using ViT as encoder, MiniLM as decoder	0.5551	IU-Xray
2023	[116]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer, CNN, RNN	ASGMD (Auxiliary signal guidance and memory-driven) network. ResNet-101 and ResNet-152 as visual feature extractors	0.489	IU-Xray
2023	[36]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	Visual Prior-based Cross-modal Alignment Network	0.489	IU-Xray
2023	[55]	Chest X-ray, CT COVID-19	MIMIC-CXR. IU-Xray. COV-CTR (728 images)	Transformer	ICT (Information calibrated transformer)	0.768	COV-CTR
2023	[16]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer, Self-Supervised Learning	S3-Net (Self-supervised dual-stream network)	0.499	IU-Xray
2023	[83]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN	TriNet (custom architecture)	0.478	IU-Xray
2023	[89]	Chest X-ray	IU-Xray. Chexpert (224,316 images)	CNN, RNN	ResNet50. CVAM+MVSL (Cross-view attention module and Medical visual-semantic LSTMs)	0.460	IU-Xray
2023	[86]	Chest X-ray	IU-Xray	CNN, RNN	Encoder-Decoder framework with UM-VES and UM-TES subnetworks and LSTM decoder	0.5881	IU-Xray
2023	[57]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	ResNet101 (visual extractor). 3-layer Transformer structure (encoder-decoder framefork). BLIP architecture	0.513	IU-Xray
2023	[56]	Chest X-ray	IU-Xray	Transformer, Contrastive Learning	MKCL (Medical knowledge with contrastive learning). ResNet-101. Transformer	0.490	IU-Xray
2023	[62]	Chest X-ray, Dermoscopy	IU-Xray, NCRC-DS (81 entities, 536 triples)	CNN, RNN, Transformer	DenseNet-121. ResNet-101. Memory-driven Transformer	0.494	IU-Xray

 doi:10.20944/preprints202511.0010.v1

2023	[115]	US (gallbladder, fetal hearth), Chest X-ray	US dataset (6,563 images and reports). Fetal Heart (FH) dataset (3,300 images and reports). MIMIC-CXR. IU-Xray	CNN, RNN	AERMNet (Attention-Enhanced Relational Memory Network)	0.890	US DATASET
2023	[59]	Chest X-ray	NIH Chest X-ray (112,120 images). MIMIC-CXR. IU-Xray	Transformer	ViT. GNN. Vector Retrieval Library. Multi-label contrastive learning. Multi-task learning	0.478	IU-Xray
2024	[17]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	Swin-Transformer	0.499	IU-Xray
2024	[51]	Chest X-ray	IU-Xray	Transformer	ViGPT2 model	0.571	IU-Xray
2024	[28]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	FMVP (Flexible multi-view paradigm)	0.499	IU-Xray
2024	[43]	Chest X-ray	Proposed Dataset (21,970 images). IU-Xray. NIH Chest X-ray	Transformer	XRaySwinGen (Swin Transformer as image encoder, GPT-2 as textual decoder)	0.731	PT BR
2024	[33]	Chest X-ray	IU-Xray	CNN, Transformer	FDT-Dr 2 T (custom famework)	0.531	IU-Xray
2024	[118]	Chest X-ray	IU-Xray. XRG-COVID-19 (8676 scans, 8676 reports)	CNN, Transformer	DSA-Transformer with ResNet-101 as the backbone	0.552	XRG-COVID-19
2024	[32]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	DenseNet-121. Transformer encoder. GPT-4	0.491	IU-Xray
2024	[58]	Oral Panoramic X-ray	Oral panoramic X-ray image-report dataset (562 sets of images and reports). MIMIC-CXR	Transformer	MLAT (Multi-Level objective Alignment Transformer)	0.5011	PAN XRAY
2024	[66]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	MRARGN (Multifocal Region-Assisted Report Generation Network)	0.502	IU-Xray
2024	[21]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	Memory-driven Transformer (based on standard Transformer architecture with relational memory added to the decoder)	0.508	IU-Xray
2024	[68]	Chest X-ray	IU-Xray	CNN, RNN, Transformer	VGG19 (CNN) pre-trained over ImageNet dataset. GloVe, fastText, ElMo, and BERT (extract textual features from the ground truth reports). Hierarchical LSTM (generate reports)	0.612	IU-Xray
2024	[93]	Chest X-ray	MIMIC-CXR. IU-Xray	RNN, Transformer, MLP	Transformer (encoder). MIX-MLP multi-label classification network. CAM (Co-attention mechanism) based on POS-SCAN. Hierarchical LSTM (decoder)	0.521	IU-Xray
2024	[45]	Chest X-ray	MIMIC-CXR	Transformer	CheXReport (Swin-B fully transformer)	0.354	MIMIC-CXR
2024	[25]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	RAMT (Relation-Aware Mean Teacher). GHFE (Graph-guided hybrid feature encoding) module. DenseNet121 (visual feature extractor). Standard Transformer (decoder)	0.482	IU-Xray
2024	[31]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	ResNet-101. Transformer (multilayer encoder and decoder)	0.514	IU-Xray
2024	[40]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	Team Role Interaction Network (TRINet)	0.445	MIMIC-CXR
2024	[26]	Polydactyly X-ray	Custom dataset (16,710 images and reports)	CNN, Transformer	Inception-V3 CNN. Transformer Architecture	0.516	PD XRAY BR
2024	[46]	Chest X-ray	MIMIC-CXR	Transformer	ViT. GPT-2 (with custom positional encoding and beam search)	*0.095 (BLEU-4)	MIMIC-CXR

 doi:10.20944/preprints202511.0010.v1

2024	[90]	Chest X-ray, Chest CT (COVID-19)	MIMIC-CXR. IU-Xray. COV-CTR	CNN, RNN	HDGAN (Hybrid Discriminator Generative Adversarial Network)	0.765	COV-CTR
2024	[117]	Chest X-ray	IU-Xray. Custom dataset (1,250 image and reports)	CNN-Transformer	CNX-B2 (CNN encoder, BioBERT transformer)	0.479	IU-Xray
2024	[37]	Chest X-ray	NIH ChestX-ray. IU-Xray	CNN, Transformer	CSAMDT (Conditional self attention memory-driven ransformer)	0.504	IU-Xray
2024	[53]	Ultrasound (gallbladder, kidney, liver), Chest X-ray	MIMIC-CXR. IU-Xray. LGK US (6,563 images and reports).	Transformer	CGFTrans (Cross-modal global feature fusion transformer)	0.684	US DATASET
2024	[52]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	TSGET (Two-stage global enhanced transformer)	0.500	IU-Xray
2024	[35]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	VCIN (Visual-textual cross-modal interaction network). ACIE (Abundant clinical information embedding). Bert-based Decoder-only Generator	0.508	IU-Xray
2024	[76]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	Memory-driven Transformer	0.539	IU-Xray
2024	[30]	Chest X-ray	MIMIC-CXR. Chest ImaGenome (237,853 images). Brown-COVID (1021 images). Penn-COVID (2879 images)	CNN, Transformer	MRANet (Multi-modality regional alignment network)	0.504	BROWN-COVID
2024	[34]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	ResNet-101. Multilayer Transformer (encoder and decoder)	0.472	IU-Xray
2024	[65]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	MeFD-Net (proposed multi expert diagnostic module). ResNet101 (visual encoder). Transformer (text generation module)	0.505	IU-Xray
2024	[19]	Chest X-ray	MIMIC-CXR. Chest ImaGenome (242,072 scene graphs)	CNN, Transformer	Faster R-CNN (object detection). GPT-2 Medium (report generation)	0.391	MIMIC-CXR
2025	[77]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	Denoising multi-level cross-attention. Contrastive learning model (with ViTs-B/16 as visual extractor, BERT as text encoder)	0.507	IU-Xray
2025	[75]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	KCAP (Knowledge-guided cross-modal alignment and progressive fusion)	0.517	IU-Xray
2025	[71]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer, ViT	ATL-CA (Adaptive topic learning and fine-grained crossmodal alignment)	0.487	IU-Xray
2025	[63]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	ADCNet (Anomaly-driven cross-modal contrastive network). ResNet-101 and Transformer encoder-decoder architecture	0.493	IU-Xray
2025	[50]	Chest X-ray	IU-Xray	Transformer	ChestX-Transcribe (combines Swin Transformer and DistilGPT)	0.675	IU-Xray
2025	[18]	Brain CT and MRI scans	RSNA-IHDC dataset (674,258 brain CT images, 19,530 patients)	CNN, Transformer	AC-BiFPN (Augmented convolutional bi-directional feature pyramid network). Transformer model	0.382	RSNA IHDC CT
2025	[61]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	DCTMN (Dual-channel transmodal memory network)	0.506	IU-Xray
2025	[74]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer, Graph reasoning network (GRN), Cross-modal Gatedgated Fusion Network (CGFN)	ResNet101 (visual feature extraction). GRN. CGFN (Cross-modal fusion network). Transformer (decoder)	0.514	IU-Xray

 doi:10.20944/preprints202511.0010.v1

2025	[47]	Spine CT	VerSe20 (300 MDCT spine images)	Transformer	ViT-Base. BioBERT BASE. MiniLM	0.7291	IU-Xray
2025	[48]	Chest X-ray	IU-Xray	Transformer	ResNet-101 with CBAM (convolutional block attention module). Cross-attention mechanism	0.456	IU-Xray
2025	[39]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	MMG (Multi-modal granularity feature fusion)	0.497	IU-Xray
2025	[24]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	RCAN (Recalibrated cross-modal alignment network)	0.521	IU-Xray
2025	[82]	Chest X-ray	IU-Xray	CNN, RNN	G-CNX (hybrid encoder–decoder architecture). ConvNeXtBase (encoder side). GRU-based RNN (decoder side)	0.6544	IU-Xray
2025	[67]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, RNN, Transformer	DPN (Dynamics priori networks) with components including DGN (Dynamic graph networks), Contrastive learning, PrKN (Prior knowledge networks). ResNet-152 (image feature extraction). SciBert (report embedding)	0.409	IU-Xray
2025	[29]	Chest X-ray, Bladder Pathology	MIMIC-CXR. IU-Xray. 4,253 bladder Pathology images.	Transformer, CNN	AHP (Adapter-enhanced hierarchical cross-modal pre-training)	0.502	IU-Xray
2025	[119]	Chest X-ray	COVIDx-CXR-2 (29,986 images). COVID-CXR (more than 900 images). BIMCV-COVID-19 (more than 10,000 images). COV-CTR. MIMIC-CXR. NIH ChestX-ray	CNN, Transformer	ResNet-50 (image encoder). BERT (text encoder). Transformer-based model (with variants using LLAMA-2-7B and Transformer-BASE, decoder)	*0.63 (BLEU-4)	COVID-19 DATASETS
2025	[20]	Chest X-ray	MIMIC-CXR. IU-Xray	CNN, Transformer	CECL (Clustering enhanced contrastive learning)	0.485	IU-Xray
2025	[69]	Chest X-ray	MIMIC-CXR. IU-Xray	Diffusion Models, RNN, CNN, Transformer	Diffusion Model-based architecture. ResNet34. Transformer structure using cross-attention	0.422	IU-Xray
2025	[60]	Chest X-ray	MIMIC-CXR. IU-Xray	Transformer	STREAM (Spatio-temporal and retrieval-augmented modelling). SwinTransformer (Swin-Base) (encoder). TinyLlama-1.1B (decoder).	0.506	IU-Xray
2025	[72]	Chest X-ray	MIMIC-CXR. ROCO (over 81,000 images)	CNN, RNN, Transformer	CAT (Cross-modal augmented transformer)	0.491	IU-Xray
2025	[38]	Chest X-ray	IU-Xray . COV-CTR	Transformer	MedVAG (Medical vision attention generation)	0.808	COV-CTR

* Indicates the **BLEU-4** value, as the study did not report the BLEU-1 metric. ** Indicates an **average of the BLEU metrics**, as the study did not report the BLEU-1 metric.

 doi:10.20944/preprints202511.0010.v1

Although these methods differ from CNN–Transformer or hybrid pipelines, they collectively illustrate a movement toward architectures that incorporate structured reasoning or external supervision to compensate for dataset and interpretability limitations. However, because their adoption remains technically experimental and narrowly scoped, their translational maturity is still preliminary relative to the more broadly validated families of models.

4. Discussion

The evolution of ARRG architectures reflects a progressive shift from sequential language models toward more expressive, attention-based frameworks capable of capturing long-range semantic dependencies. Early CNN–RNN pipelines demonstrated the feasibility of translating image-derived representations into coherent textual descriptions, but their reliance on stepwise decoding limited their ability to capture global contextual dependencies within radiological narratives [95,96]. Subsequent extensions attempted to mitigate these constraints through multi-task learning and co-attention mechanisms, which improved alignment between visual features and semantic structure but remained fundamentally restricted by the sequential nature of RNN-based decoding [97]. The introduction of Transformer-based models marked a methodological inflection point by enabling parallelized processing, improved feature integration, and richer contextual reasoning [95]. Hybrid architectures further enhanced performance by combining CNNs for localized feature extraction with Transformers for global semantic modeling, while more recent developments incorporate memory augmentation, medical priors, or retrieval-based alignment mechanisms to compensate for limited contextual cues in public datasets.

More recent research has explored advanced strategies to further improve report quality and strengthen alignment with radiological reasoning. Memory-augmented architectures and models that incorporate structured medical knowledge have shown promising performance, particularly on large public benchmarks such as MIMIC-CXR [96]. These systems typically enrich representation space through the integration of pre-built knowledge graphs or retrieval-based mechanisms that draw from similar reports or pathology patterns [49,98,99], moving closer to the way radiologists ground their interpretation in prior clinical context. However, although these mechanisms improve semantic alignment, they do not yet guarantee diagnostic accountability or case-level reasoning, limiting their contribution to true clinical readiness.

Other innovations include Region-guided Report Generation (RGRG), which enhances explainability by anchoring narrative content to localized anatomical regions [100]; RECAP, which introduces temporal reasoning to capture disease progression and longitudinal consistency [101]; and UAR (“unify, align, and refine”), a framework designed to align visual and textual features across multiple semantic levels [102]. Progressive-generation strategies have also been proposed to iteratively refine report outputs, leading to more stable, coherent, and clinically focused narratives [103]. These advances indicate that recent improvements in ARRG extend beyond raw language-modeling capacity to increasingly incorporate mechanisms that emulate elements of human diagnostic reasoning.

Collectively, this architectural trajectory demonstrates substantial technical maturation, although increased model complexity has not yet translated into consistent improvements in clinically grounded interpretability or translational maturity.

However, evaluating ARRG systems remains one of the most critical challenges in the field. Widely used NLG metrics such as BLEU [104] and ROUGE [105] primarily measure surface-level similarity based on n-gram overlap [105–107]. While useful in some contexts, these metrics often fail to capture deeper semantic equivalence, paraphrastic variation, or clinically relevant word ordering. In fact, studies have shown that BLEU correlates poorly with human judgment in image captioning tasks [106,107], and its alignment with expert radiologist evaluations for CXR reports is particularly weak [108].

Another limitation lies in the availability of reference reports. It is estimated that around 50 human-written reports per image are needed to achieve reliable consensus, yet most public datasets

 doi:10.20944/preprints202511.0010.v1

provide only five [107]. Overcoming these limitations will require not only more sophisticated model architectures but also the development of large, high-quality datasets, such as MIMIC-CXR, combined with clinically aligned evaluation metrics like RadGraph F1 and RadCliQ [108], which better reflect the true diagnostic quality and clinical usefulness of generated reports.

Medical databases often demonstrate a tenuous connection to authentic clinical scenarios. The process of capturing real-world medical data is fraught with challenges, leading to datasets that are limited and biased towards common cases while marginalizing critical abnormalities. Such limitations restrict linguistic diversity and impede the development of varied descriptions, particularly for rare and nuanced cases, which are crucial for precise clinical diagnosis. Moreover, these biases not only constrain linguistic variability but also undermine the generalizability of models across different institutions and clinical contexts, ultimately reducing their applicability in diverse real-world settings. As a result, models dependent on these datasets may experience deficiencies in accuracy and reliability when deployed in clinical practice.

Compared to previous reviews, this work provides a broader and more up-to-date overview of automated radiology report generation. While Kaur et al. [109] focused exclusively on CXR, limiting generalization to other modalities, our review highlights the need to extend ARRG research beyond thoracic imaging. Similarly, the review by Monshi et al. [110] emphasized early CNN-RNN approaches but does not cover recent advances such as Transformer-based models and knowledge-enhanced frameworks. Furthermore, although Liao et al. [111] provided a systematic analysis of datasets and evaluation methods, their discussion lacks a strong connection to clinical challenges and real-world applicability. In contrast, this review not only synthesizes current technical trends but also situates them within the broader clinical workflow, emphasizing integration into diagnostic practice, highlighting limitations of existing evaluation metrics, and proposing future research directions aimed at improving both the accuracy and practical utility of ARRG systems.

This review provides a comprehensive and up-to-date overview of the current state of ARRG using DL, with a particular focus on architectural trends, evaluation practices, and clinical applications. Additionally, by capturing not only technical details but also clinical context, this work contributes to bridging the gap between algorithmic development and real-world diagnostic needs. Nevertheless, some limitations should be acknowledged. The review predominantly reflects research efforts focused on CXR, largely influenced by the accessibility of public datasets like MIMIC-CXR and the urgency created by the COVID-19 pandemic. Consequently, other anatomical regions and imaging modalities remain underexplored, highlighting the need for broader dataset development and more diverse applications. Furthermore, while the review describes prevailing evaluation metrics, it also reveals the ongoing limitations of these measures in capturing true clinical relevance. Looking forward, future research should prioritize clinically aligned evaluation frameworks, expand model development beyond thoracic imaging, and explore the integration of large language models and domain-specific knowledge to improve both report quality and diagnostic accuracy. Addressing these gaps is essential to realizing the full potential of automated report generation in supporting radiologists and enhancing healthcare delivery.

5. Conclusions

This systematic review synthesizes the current state of ARRG using DL, highlighting both its methodological evolution and its emerging clinical relevance. The key findings can be summarized as follows:

- The field remains heavily concentrated on chest radiography, with more than 87% of studies based on CXR datasets. This reflects public data availability and the acceleration of thoracic imaging research during the COVID-19 pandemic, but also exposes a lack of anatomical diversity that limits generalizability to other diagnostic domains encountered in routine radiological practice.
- Hybrid architectures, particularly CNN–Transformer combinations, represent the dominant methodological trend (73% of included studies). By leveraging CNNs for localized visual

 doi:10.20944/preprints202511.0010.v1

encoding and Transformer modules for contextual reasoning, these models generate reports with greater coherence and abnormality representation, reducing variability and supporting more consistent documentation.

- The increased use of memory modules, medical knowledge graphs, and cross-modal alignment mechanisms demonstrates a clear shift toward clinically informed modeling. These strategies improve factual grounding by embedding structured domain knowledge into the generation process and aligning outputs more closely with expert reasoning.
- However, current evaluation frameworks remain poorly aligned with clinical decision-making. Metrics such as BLEU and ROUGE capture surface-level similarity but do not reflect diagnostic adequacy or patient management utility, underscoring the need for evaluation standards that measure whether generated reports truly support radiological interpretation and workflow reliability.

Overall, ARRG has achieved meaningful technical progress, yet its translation into real clinical environments remains constrained by limited anatomical coverage, shallow evaluation standards, and insufficient external validation. For these systems to evolve from experimental prototypes into trustworthy decision support tools, future research must prioritize clinically grounded benchmarking, greater dataset diversity, and integration pathways that reflect the realities of radiological practice. As these gaps are progressively addressed, ARRG has the potential to become a scalable and clinically accountable complement to radiological reporting, provided that future developments successfully bridge the remaining gap in clinical readiness.

Author Contributions: Conceptualization, P.M.-R., J.J.-R., M.F.V.-D., M.P. and A.V.-B.; methodology, P.M.-R., J.J.-R., M.F.V.-D., M.P. and A.V.-B.; validation, P.M.-R., J.J.-R. and M.F.V.-D.; formal analysis P.M.-R., J.J.-R., M.F.V.-D., M.P. and A.V.-B.; investigation, P.M.-R., J.J.-R. and M.F.V.-D.; writing—original draft preparation, P.M.-R.; writing—review and editing, P.M.-R., J.J.-R., M.F.V.-D., M.P. and A.V.-B.; project administration, P.M.-R.; funding acquisition, P.M.-R.. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Agencia Nacional de Investigación y Desarrollo (ANID), Chile, through the *Doctorado en Chile Scholarship Program, Academic Year 2025 (Grant No. 1340/2025)*.

Data Availability Statement: All data supporting the findings of this systematic review are derived from previously published articles that are publicly available in their corresponding databases (PubMed, Scopus, and WoS). As this study did not generate or analyze primary datasets, no new data was created. The full list of publications included in this review is provided in the manuscript and serves as a direct reference to the original data sources.

Conflicts of Interest: The authors declare no financial or non-financial conflicts of interest that could be perceived as influencing the work reported in this manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

ARRG	Automatic Radiology Report Generation
DL	Deep learning
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
COVID-19	Coronavirus disease 2019
CNNs	Convolutional neural networks
RNNs	Recurrent neural networks
LSTM	Long short-term memory
CLIP	Contrastive language-image pretraining
LLMs	Large language models
MIMIC	Medical information mart for intensive care database
MIMIC-CXR	MIMIC-Chest X-ray
IU-Xray	Indiana university chest X-ray collection

 doi:10.20944/preprints202511.0010.v1

BLEU	Bilingual evaluation understudy
ROUGE	Recall-oriented understudy for gisting evaluation
BERT	Bidirectional encoder representations from transformers
AUC	Area under the curve
IEEE	Institute of electrical and electronics engineers
ACM	Association for computing machinery
WoS	Web of science
GNNs	Graph neural networks
GAT	Graph attention network
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
CXR	Chest X-ray
CT	Computed tomography
MRI	Magnetic resonance imaging
CTN	Contrastive triplet network
SOTA	State of the art
METEOR	Metric for evaluation of translation with explicit ordering
CIDEr	Consensus-based image description evaluation
ViT	Vision transformers
GPT	Generative pre-trained transformers
NLG	Natural language generation
GRU	Gated recurrent unit
SVMs	Support vector machines
KG	Knowledge graph
DS	Dataset
US	Ultrasound
DBT	Digital breast tomosynthesis
SFNet	Semantic fusion network
CDGPT	Conditioned distil generative pre-trained transformer
MLTL	Multi level transfer learning
HReMRG	Hybrid reinforced medical report generation method
ATAG	Attributed abnormality graph
AMLMA	Adaptive multilevel multi-attention
VTI	Variational topic inference
CAMANet	Class activation map guided attention network
MFOT	Multi-feature optimization transformer
TrMRG	Transformer medical report generator
ASGMD	Auxiliary signal guidance and memory-driven
ICT	Information calibrated transformer
CVAM	Cross-view attention module
MVSL	Medical visual-semantic LSTMs
MKCL	Medical knowledge with contrastive learning
AERMNet	Attention-Enhanced Relational Memory Network
FMVP	Flexible multi-view paradigm
RAMT	Relation-aware mean teacher
GHFE	Graph-guided hybrid feature encoding
CSAMDT	Conditional self attention memory-driven transformer
CGFTrans	Cross-modal global feature fusion transformer
TSGET	Two-stage global enhanced transformer
VCIN	Visual-textual cross-modal interaction network
ACIE	Abundant clinical information embedding
MRANet	Multi-modality regional alignment network
KCAP	Knowledge-guided cross-modal alignment and progressive fusion
ATL-CA	Adaptive topic learning and fine-grained crossmodal alignment
ADCNet	Anomaly-driven cross-modal contrastive network
AC-BiFPN	Augmented convolutional bi-directional feature pyramid network
DCTMN	Dual-channel transmodal memory network
GRN	Graph reasoning network
CGFN	Cross-modal gated fusion network
CBAM	Convolutional block attention module

 doi:10.20944/preprints202511.0010.v1

MMG	Multi-modal granularity feature fusion
RCAN	Recalibrated cross-modal alignment network
DPN	Dynamics priori networks
DGN	Dynamic graph networks
PrKN	Prior knowledge networks
AHP	Adapter-enhanced hierarchical cross-modal pre-training
CECL	Clustering enhanced contrastive learning
STREAM	Spatio-temporal and retrieval-augmented modelling
CAT	Cross-modal augmented transformer
MedVAG	Medical vision attention generation
RGRG	Region-guided report generation
UAR	Unify, align and refine

References

1. Ramirez-Alonso, G.; Prieto-Ordaz, O.; López-Santillan, R.; Montes-Y-Gómez, M. Medical report generation through radiology images: An overview. *IEEE Lat. Am. Trans.* **2022**, *20*(6), 986–999. https://doi.org/10.1109/TLA.2022.9757742
2. Kaur, N.; Mittal, A.; Singh, G. Methods for automatic generation of radiological reports of chest radiographs: A comprehensive survey. *Multimed. Tools Appl.* **2022**, *81*(10), 13409–13439. https://doi.org/10.1007/s11042-021-11272-6
3. Pang, T.; Li, P.; Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. *Biomed. Eng. Online* **2023**, *22*(1), 48. https://doi.org/10.1186/s12938-023-01113-y
4. Azad, R.; Kazerouni, A.; Heidari, M.; Khodapanah Aghdam, E.; Molaei, A.; Jia, Y.; et al. Advances in medical image analysis with vision transformers: A comprehensive review. *Med. Image Anal.* **2024**, *91*, 103000. https://doi.org/10.1016/j.media.2023.103000
5. Sun, Z.; Lin, M.; Zhu, Q.; Xie, Q.; Wang, F.; Lu, Z.; et al. A scoping review on multimodal deep learning in biomedical images and texts. *J. Biomed. Inform.* **2023**, *146*, 104482. https://doi.org/10.1016/j.jbi.2023.104482
6. Sloan, P.; Clatworthy, P.L.; Simpson, E.; Mirmehdi, M. Automated radiology report generation: A review of recent advances. *IEEE Rev. Biomed. Eng.* **2025**, *18*, 368–387. https://doi.org/10.1109/RBME.2024.3408456
7. Guo, L.; Tahir, A.M.; Zhang, D.; Wang, Z.J.; Ward, R.K. Automatic medical report generation: Methods and applications. *APSIPA Trans. Signal Inf. Process.* **2024**, *13*(1). https://doi.org/10.1561/116.20240044
8. Shen, Y.; Xu, Y.; Ma, J.; Rui, W.; Zhao, C.; Heacock, L.; et al. Multi-modal large language models in radiology: Principles, applications, and potential. *Abdom. Radiol.* **2024**. https://doi.org/10.1007/s00261-024-04708-8
9. Nakaura, T.; Ito, R.; Ueda, D.; Nozaki, T.; Fushimi, Y.; Matsui, Y.; et al. The impact of large language models on radiology: A guide for radiologists on the latest innovations in AI. *Jpn. J. Radiol.* **2024**, *42*(7), 685–696. https://doi.org/10.1007/s11604-024-01552-0
10. Nerella, S.; Bandyopadhyay, S.; Zhang, J.; Contreras, M.; Siegel, S.; Bumin, A.; et al. Transformers and large language models in healthcare: A review. *Artif. Intell. Med.* **2024**, *154*, 102900. https://doi.org/10.1016/j.artmed.2024.102900
11. Ouis, M.Y.; Akhloufi, M.A. Deep learning for report generation on chest X-ray images. *Comput. Med. Imaging Graph.* **2024**, *111*, 102320. https://doi.org/10.1016/j.compmedimag.2023.102320
12. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. https://doi.org/10.1038/s41591-024-03425-5
13. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. https://doi.org/10.1136/bmj.n160
14. Yang, Y.; Yu, J.; Jiang, H.; Han, W.; Zhang, J.; Jiang, W. A contrastive triplet network for automatic chest X-ray reporting. *Neurocomputing* **2022**, *502*, 71–83. https://doi.org/10.1016/j.neucom.2022.06.063
15. Nicolson, A.; Dowling, J.; Koopman, B. Improving chest X-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **2023**, *144*, 102633. https://doi.org/10.1016/j.artmed.2023.102633

16. Pan, R.; Ran, R.; Hu, W.; Zhang, W.; Qin, Q.; Cui, S. S3-Net: A self-supervised dual-stream network for radiology report generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*(3), 1448–1459. https://doi.org/10.1109/JBHI.2023.3345932
17. Pan, Y.; Liu, L.J.; Yang, X.B.; Peng, W.; Huang, Q.S. Chest radiology report generation based on cross-modal multi-scale feature fusion. *J. Radiat. Res. Appl. Sci.* **2024**, *17*(1), 100823. https://doi.org/10.1016/j.jrras.2024.100823
18. Bouslimi, R.; Trabelsi, H.; Karaa, W.B.A.; Hedhli, H. AI-driven radiology report generation for traumatic brain injuries. *J. Imaging Inform. Med.* **2025**. https://doi.org/10.1007/s10278-025-01411-y
19. Zhang, K.; Yang, Y.; Yu, J.; Fan, J.; Jiang, H.; Huang, Q.; et al. Attribute prototype-guided iterative scene graph for explainable radiology report generation. *IEEE Trans. Med. Imaging* **2024**, *43*(12), 4470–4482. https://doi.org/10.1109/TMI.2024.3424505
20. Liu, X.; Xin, J.; Shen, Q.; Li, C.; Huang, Z.; Wang, Z. End-to-end clustering enhanced contrastive learning for radiology reports generation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*(2), 1780–1794. https://doi.org/10.1109/TETCI.2024.3449876
21. Liu, X.; Xin, J.; Dai, B.; Shen, Q.; Huang, Z.; Wang, Z. Label correlated contrastive learning for medical report generation. *Comput. Methods Programs Biomed.* **2025**, *258*, 108482. https://doi.org/10.1016/j.cmpb.2024.108482
22. Shang, C.; Cui, S.; Li, T.; Wang, X.; Li, Y.; Jiang, J. MATNet: Exploiting multi-modal features for radiology report generation. *IEEE Signal Process. Lett.* **2022**, *29*, 2692–2696. https://doi.org/10.1109/LSP.2022.3229844
23. Yan, B.; Pei, M.; Zhao, M.; Shan, C.; Tian, Z. Prior guided transformer for accurate radiology reports generation. *IEEE J. Biomed. Health Inform.* **2022**, *26*(11), 5631–5640. https://doi.org/10.1109/JBHI.2022.3197162
24. Hou, X.; Li, X.; Liu, Z.; Sang, S.; Lu, M.; Zhang, Y. Recalibrated cross-modal alignment network for radiology report generation with weakly supervised contrastive learning. *Expert Syst. Appl.* **2025**, *269*, 126394. https://doi.org/10.1016/j.eswa.2025.126394
25. Zhang, K.; Jiang, H.; Zhang, J.; Fan, J.; Yu, J.; et al. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Trans. Multimedia* **2024**, *26*, 904–915. https://doi.org/10.1109/TMM.2023.3273390
26. Vieira, P.D.A.; Mathew, M.J.; Santos Neto, P.D.A.D.; Silva, R.R.V.E. The automated generation of medical reports from polydactyly X-ray images using CNNs and transformers. *Appl. Sci.* **2024**, *14*(15), 6566. https://doi.org/10.3390/app14156566
27. Wang, J.; Bhalerao, A.; Yin, T.; See, S.; He, Y. CAMANet: Class activation map guided attention network for radiology report generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*(4), 2199–2210. https://doi.org/10.1109/JBHI.2024.3354712
28. Liu, Z.; Zhu, Z.; Zheng, S.; Zhao, Y.; He, K.; Zhao, Y. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Trans. Multimedia* **2024**, *26*, 5987–5995. https://doi.org/10.1109/TMM.2023.3342691
29. Yu, T.; Lu, W.; Yang, Y.; Han, W.; Huang, Q.; Yu, J.; et al. Adapter-enhanced hierarchical cross-modal pre-training for lightweight medical report generation. *IEEE J. Biomed. Health Inform.* **2025**, pages 1–15. https://doi.org/10.1109/JBHI.2025.3535699
30. Zhong, Z.; Li, J.; Sollee, J.; Collins, S.; Bai, H.; Zhang, P.; et al. Multi-modality regional alignment network for COVID X-ray survival prediction and report generation. *IEEE J. Biomed. Health Inform.* **2024**, pages 1–11. https://doi.org/10.1109/JBHI.2024.3417849
31. Li, S.; Qiao, P.; Wang, L.; Ning, M.; Yuan, L.; Zheng, Y.; et al. An organ-aware diagnosis framework for radiology report generation. *IEEE Trans. Med. Imaging* **2024**, *43*(12), 4253–4265. https://doi.org/10.1109/TMI.2024.3421599
32. Li, H.; Wang, H.; Sun, X.; He, H.; Feng, J. Context-enhanced framework for medical image report generation using multimodal contexts. *Knowl.-Based Syst.* **2025**, *310*, 112913. https://doi.org/10.1016/j.knosys.2024.112913
33. Sharma, D.; Dhiman, C.; Kumar, D. FDT–Dr2T: A unified dense radiology report generation transformer framework for X-ray images. *Mach. Vis. Appl.* **2024**, *35*(4), 68. https://doi.org/10.1007/s00138-024-01544-0

 doi:10.20944/preprints202511.0010.v1

34. Liu, A.; Guo, Y.; Yong, J.H.; Xu, F. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Trans. Med. Imaging* **2024**, *43*(7), 2657–2669. https://doi.org/10.1109/TMI.2024.3372638
35. Zhang, W.; Cai, B.; Hu, J.; Qin, Q.; Xie, K. Visual-textual cross-modal interaction network for radiology report generation. *IEEE Signal Process. Lett.* **2024**, *31*, 984–988. https://doi.org/10.1109/LSP.2024.3379005
36. Zhang, S.; Zhou, C.; Chen, L.; Li, Z.; Gao, Y.; Chen, Y. Visual prior-based cross-modal alignment network for radiology report generation. *Comput. Biol. Med.* **2023**, *166*, 107522. https://doi.org/10.1016/j.compbiomed.2023.107522
37. Shahzadi, I.; Madni, T.M.; Janjua, U.I.; Batool, G.; Naz, B.; Ali, M.Q.; et al. CSAMDT: Conditional self-attention memory-driven transformers for radiology report generation from chest X-ray. *J. Imaging Inform. Med.* **2024**, *37*(6), 2825–2837. https://doi.org/10.1007/s10278-024-01126-6
38. Varol Arısoy, M.; Arısoy, A.; Uysal, I. A vision attention driven language framework for medical report generation. *Sci. Rep.* **2025**, *15*(1), 10704. https://doi.org/10.1038/s41598-025-95666-8
39. Fang, J.; Xing, S.; Li, K.; Guo, Z.; Li, G.; Yu, C. Automated generation of chest X-ray imaging diagnostic reports by multimodal and multi-granularity features fusion. *Biomed. Signal Process. Control* **2025**, *105*, 107562. https://doi.org/10.1016/j.bspc.2025.107562
40. Zheng, Z.; Zhang, Y.; Liang, E.; Weng, Z.; Chai, J.; Li, J. TRINet: Team role interaction network for automatic radiology report generation. *Comput. Biol. Med.* **2024**, *183*, 109275. https://doi.org/10.1016/j.compbiomed.2024.109275
41. Vendrow, E.; Schonfeld, E. Understanding transfer learning for chest radiograph clinical report generation with modified transformer architectures. *Heliyon* **2023**, *9*(7), e17968. https://doi.org/10.1016/j.heliyon.2023.e17968
42. Mohsan, M.M.; Akram, M.U.; Rasool, G.; Alghamdi, N.S.; Baqai, M.A.A.; Abbas, M. Vision transformer and language model-based radiology report generation. *IEEE Access* **2023**, *11*, 1814–1824. https://doi.org/10.1109/ACCESS.2022.3232719
43. Veras Magalhaes, G.; De S. Santos, R.L.; Vogado, L.H.S.; Cardoso De Paiva, A.; De Alcantara Dos Santos Neto, P. XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model. *Heliyon* **2024**, *10*(7), e27516. https://doi.org/10.1016/j.heliyon.2024.e27516
44. Wang, Z.; Han, H.; Wang, L.; Li, X.; Zhou, L. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Trans. Med. Imaging* **2022**, *41*(10), 2803–2813. https://doi.org/10.1109/TMI.2022.3171661
45. Zeiser, F.A.; Da Costa, C.A.; De Oliveira Ramos, G.; Maier, A.; Da Rosa Righi, R. CheXReport: A transformer-based architecture to generate chest X-ray reports suggestions. *Expert Syst. Appl.* **2024**, *255*, 124644. https://doi.org/10.1016/j.eswa.2024.124644
46. Leonardi, G.; Portinale, L.; Santomauro, A. Enhancing radiology report generation through pre-trained language models. *Prog. Artif. Intell.* **2024**. https://doi.org/10.1007/s13748-024-00358-5
47. Batool, H.; Mukhtar, A.; Khawaja, S.G.; Alghamdi, N.S.; Khan, A.M.; Qayyum, A.; et al. Knowledge distillation and transformer-based framework for automatic spine CT report generation. *IEEE Access* **2025**, *13*, 42949–42964. https://doi.org/10.1109/ACCESS.2025.3546131
48. Zhao, J.; Yao, W.; Sun, L.; Shi, L.; Kuang, Z.; Wu, C.; et al. Automated chest X-ray diagnosis report generation with cross-attention mechanism. *Appl. Sci.* **2025**, *15*(1), 343. https://doi.org/10.3390/app15010343
49. Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; Fahmy, A. Automated radiology report generation using conditioned transformers. *Informatics Med. Unlocked* **2021**, *24*, 100557. https://doi.org/10.1016/j.imu.2021.100557
50. Singh, P.; Singh, S. ChestX-Transcribe: A multimodal transformer for automated radiology report generation from chest X-rays. *Front. Digit. Health* **2025**, *7*, 1535168. https://doi.org/10.3389/fdgth.2025.1535168
51. Raminedi, S.; Shridevi, S.; Won, D. Multi-modal transformer architecture for medical image analysis and automated report generation. *Sci. Rep.* **2024**, *14*(1), 19281. https://doi.org/10.1038/s41598-024-69981-5

 doi:10.20944/preprints202511.0010.v1

52. Yi, X.; Fu, Y.; Liu, R.; Hu, Y.; Zhang, H.; Hua, R. TSGET: Two-stage global enhanced transformer for automatic radiology report generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*(4), 2152–2162. https://doi.org/10.1109/JBHI.2024.3350077
53. Xu, L.; Tang, Q.; Zheng, B.; Lv, J.; Li, W.; Zeng, X. CGFTrans: Cross-modal global feature fusion transformer for medical report generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*(9), 5600–5612. https://doi.org/10.1109/JBHI.2024.3414413
54. Wang, R.; Hua, R. Generating radiology reports via multi-feature optimization transformer. *KSII Trans. Internet Inf. Syst.* **2023**, *17*(10). https://doi.org/10.3837/tiis.2023.10.010
55. Zhang, J.; Shen, X.; Wan, S.; Goudos, S.K.; Wu, J.; Cheng, M.; et al. A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE J. Biomed. Health Inform.* **2023**, *27*(10), 5110–5121. https://doi.org/10.1109/JBHI.2023.3236661
56. Hou, X.; Liu, Z.; Li, X.; Li, X.; Sang, S.; Zhang, Y. MKCL: Medical knowledge with contrastive learning model for radiology report generation. *J. Biomed. Inform.* **2023**, *146*, 104496. https://doi.org/10.1016/j.jbi.2023.104496
57. Zhao, G.; Zhao, Z.; Gong, W.; Li, F. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. *Artif. Intell. Med.* **2023**, *146*, 102714. https://doi.org/10.1016/j.artmed.2023.102714
58. Gao, N.; Yao, R.; Liang, R.; Chen, P.; Liu, T.; Dang, Y. Multi-level objective alignment transformer for fine-grained oral panoramic X-ray report generation. *IEEE Trans. Multimedia* **2024**, *26*, 7462–7474. https://doi.org/10.1109/TMM.2024.3368922
59. Guo, K.; Zheng, S.; Huang, R.; Gao, R. Multi-task learning for lung disease classification and report generation via prior graph structure and contrastive learning. *IEEE Access* **2023**, *11*, 110888–110898. https://doi.org/10.1109/ACCESS.2023.3322425
60. Yang, Y.; You, X.; Zhang, K.; Fu, Z.; Wang, X.; Ding, J.; et al. Spatio-temporal and retrieval-augmented modelling for chest X-ray report generation. *IEEE Trans. Med. Imaging* **2025**, pages 1–1. https://doi.org/10.1109/TMI.2025.3554498
61. Dong, Z.; Lian, J.; Zhang, X.; Zhang, B.; Liu, J.; Zhang, J.; et al. A chest imaging diagnosis report generation method based on dual-channel transmodal memory network. *Biomed. Signal Process. Control* **2025**, *100*, 107021. https://doi.org/10.1016/j.bspc.2024.107021
62. Xu, D.; Zhu, H.; Huang, Y.; Jin, Z.; Ding, W.; Li, H.; et al. Vision-knowledge fusion model for multi-domain medical report generation. *Inf. Fusion* **2023**, *97*, 101817. https://doi.org/10.1016/j.inffus.2023.101817
63. Liu, Y.; Zhang, J.; Liu, K.; Tan, L. ADCNet: Anomaly-driven cross-modal contrastive network for medical report generation. *Electronics* **2025**, *14*(3), 532. https://doi.org/10.3390/electronics14030532
64. Yan, S.; Cheung, W.K.; Chiu, K.; Tong, T.M.; Cheung, K.C.; See, S. Attributed abnormality graph embedding for clinically accurate X-ray report generation. *IEEE Trans. Med. Imaging* **2023**, *42*(8), 2211–2222. https://doi.org/10.1109/TMI.2023.3245608
65. Ran, R.; Pan, R.; Yang, W.; Deng, Y.; Zhang, W.; Hu, W.; et al. MeFD-Net: Multi-expert fusion diagnostic network for generating radiology image reports. *Appl. Intell.* **2024**, *54*(22), 11484–11495. https://doi.org/10.1007/s10489-024-05680-y
66. Yang, B.; Lei, H.; Huang, H.; Han, X.; Cai, Y. DPN: Dynamics priori networks for radiology report generation. *Tsinghua Sci. Technol.* **2025**, *30*(2), 600–609. https://doi.org/10.26599/TST.2023.9010134
67. Yang, B.; Lei, H.; Huang, H.; Han, X.; Cai, Y. DPN: Dynamics priori networks for radiology report generation. *Tsinghua Sci. Technol.* **2025**, *30*(2), 600–609. https://doi.org/10.26599/TST.2023.9010134
68. Alotaibi, F.S.; Kaur, N. Radiological report generation from chest X-ray images using pre-trained word embeddings. *Wirel. Pers. Commun.* **2023**, *133*(4), 2525–2540. https://doi.org/10.1007/s11277-024-10886-x
69. Sun, S.; Su, Z.; Meizhou, J.; Feng, Y.; Hu, Q.; Luo, J.; et al. Optimizing medical image report generation through a discrete diffusion framework. *J. Supercomput.* **2025**, *81*(5), 637. https://doi.org/10.1007/s11227-025-07111-2
70. Kaur, N.; Mittal, A. RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings. *J. Biomed. Inform.* **2022**, *135*, 104220. https://doi.org/10.1016/j.jbi.2022.104220

 doi:10.20944/preprints202511.0010.v1

71. Mei, X.; Yang, L.; Gao, D.; Cai, X.; Han, J.; Liu, T. Adaptive medical topic learning for enhanced fine-grained cross-modal alignment in medical report generation. *IEEE Trans. Multimedia* **2025**, pages 1–12. https://doi.org/10.1109/TMM.2025.3543101
72. Tang, Y.; Yuan, Y.; Tao, F.; Tang, M. Cross-modal augmented transformer for automated medical report generation. *IEEE J. Transl. Eng. Health Med.* **2025**, *13*, 33–48. https://doi.org/10.1109/JTEHM.2025.3536441
73. Najdenkoska, I.; Zhen, X.; Worring, M.; Shao, L. Uncertainty-aware report generation for chest X-rays by variational topic inference. *Med. Image Anal.* **2022**, *82*, 102603. https://doi.org/10.1016/j.media.2022.102603
74. Zhang, J.; Cheng, M.; Li, X.; Shen, X.; Wan, Y.; Zhu, J.; et al. Generating medical report via joint probability graph reasoning. *Tsinghua Sci. Technol.* **2025**, *30*(4), 1685–1699. https://doi.org/10.26599/TST.2024.9010058
75. Huang, L.; Cao, Y.; Jia, P.; Li, C.; Tang, J.; Li, C. Knowledge-guided cross-modal alignment and progressive fusion for chest X-ray report generation. *IEEE Trans. Multimedia* **2025**, *27*, 557–567. https://doi.org/10.1109/TMM.2024.3521728
76. Yi, X.; Fu, Y.; Yu, J.; Liu, R.; Zhang, H.; Hua, R. LHR-RFL: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE Trans. Med. Imaging* **2025**, *44*(3), 1494–1504. https://doi.org/10.1109/TMI.2024.3507073
77. Zhu, D.; Liu, L.; Yang, X.; Liu, L.; Peng, W. Denoising multi-level cross-attention and contrastive learning for chest radiology report generation. *J. Imaging Inform. Med.* **2025**. https://doi.org/10.1007/s10278-025-01422-9
78. Li, H.; Liu, X.; Jia, D.; Chen, Y.; Hou, P.; Li, H. Research on chest radiography recognition model based on deep learning. *Math. Biosci. Eng.* **2022**, *19*(11), 11768–11781. https://doi.org/10.3934/mbe.2022548
79. Sirshar, M.; Paracha, M.F.K.; Akram, M.U.; Alghamdi, N.S.; Zaidi, S.Z.Y.; Fatima, T. Attention-based automated radiology report generation using CNN and LSTM. *PLOS ONE* **2022**, *17*(1), e0262209. https://doi.org/10.1371/journal.pone.0262209
80. Hou, D.; Zhao, Z.; Liu, Y.; Chang, F.; Hu, S. Automatic report generation for chest X-ray images via adversarial reinforcement learning. *IEEE Access* **2021**, *9*, 21236–21250. https://doi.org/10.1109/ACCESS.2021.3056175
81. Kaur, N.; Mittal, A. CADxReport: Chest X-ray report generation using co-attention mechanism and reinforcement learning. *Comput. Biol. Med.* **2022**, *145*, 105498. https://doi.org/10.1016/j.compbiomed.2022.105498
82. Ucan, M.; Kaya, B.; Kaya, M. Generating medical reports with a novel deep learning architecture. *Int. J. Imaging Syst. Technol.* **2025**, *35*(2), e70062. https://doi.org/10.1002/ima.70062
83. Yang, Y.; Yu, J.; Zhang, J.; Han, W.; Jiang, H.; Huang, Q. Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Trans. Multimedia* **2023**, *25*, 167–178. https://doi.org/10.1109/TMM.2021.3122542
84. Gajbhiye, G.O.; Nandedkar, A.V.; Faye, I. Translating medical image to radiological report: Adaptive multilevel multi-attention approach. *Comput. Methods Programs Biomed.* **2022**, *221*, 106853. https://doi.org/10.1016/j.cmpb.2022.106853
85. Kaur, N.; Mittal, A. CheXPrune: Sparse chest X-ray report generation model using multi-attention and one-shot global pruning. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*(6), 7485–7497. https://doi.org/10.1007/s12652-022-04454-z
86. Shetty, S.; Ananthanarayana, V.S.; Mahale, A. Cross-modal deep learning-based clinical recommendation system for radiology report generation from chest X-rays. *Int. J. Eng.* **2023**, *36*(8), 1569–1577. https://doi.org/10.5829/IJE.2023.36.08B.16
87. Zeng, X.; Wen, L.; Xu, Y.; Ji, C. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. *Comput. Methods Programs Biomed.* **2020**, *197*, 105700. https://doi.org/10.1016/j.cmpb.2020.105700
88. Xu, Z.; Xu, W.; Wang, R.; Chen, J.; Qi, C.; Lukasiewicz, T. Hybrid reinforced medical report generation with M-linear attention and repetition penalty. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*(2), 2206–2222. https://doi.org/10.1109/TNNLS.2023.3343391

89. Gu, Y.; Li, R.; Wang, X.; Zhou, Z. Automatic medical report generation based on cross-view attention and visual-semantic long short-term memorys. *Bioengineering* **2023**, *10*(8), 966. https://doi.org/10.3390/bioengineering10080966
90. Zhang, J.; Cheng, M.; Cheng, Q.; Shen, X.; Wan, Y.; Zhu, J.; et al. Hierarchical medical image report adversarial generation with hybrid discriminator. Artif. Intell. Med. **2024**, 151, 102846. https://doi.org/10.1016/j.artmed.2024.102846
91. Paalvast, O.; Nauta, M.; Koelle, M.; Geerdink, J.; Vijlbrief, O.; Hegeman, J.H.; et al. Radiology report generation for proximal femur fractures using deep classification and language generation models. *Artif. Intell. Med.* **2022**, *128*, 102281. https://doi.org/10.1016/j.artmed.2022.102281
92. Loveymi, S.; Dezfoulian, M.H.; Mansoorizadeh, M. Automatic generation of structured radiology reports for volumetric computed tomography images using question-specific deep feature extraction and learning. *J. Med. Signals Sens.* **2021**, *11*(3), 194–207. https://doi.org/10.4103/jmss.JMSS_21_20
93. Sun, S.; Mei, Z.; Li, X.; Tang, T.; Li, Z.; Wu, Y. A label information fused medical image report generation framework. *Artif. Intell. Med.* **2024**, *150*, 102823. https://doi.org/10.1016/j.artmed.2024.102823
94. Zhang, D.; Ren, A.; Liang, J.; Liu, Q.; Wang, H.; Ma, Y. Improving medical X-ray report generation by using knowledge graph. *Appl. Sci.* **2022**, *12*(21), 11111. https://doi.org/10.3390/app122111111
95. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*; 2017.
96. 96. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating radiology reports via memory-driven transformer. arXiv preprint **2022**, arXiv:2010.16056. https://github.com/zhjohnchan/R2Gen
97. Jing, B.; Xie, P.; Xing, E.P. On the automatic generation of medical imaging reports. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2018.
98. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021. https://doi.org/10.1109/CVPR46437.2021.01354
99. Yang, S.; Wu, X.; Ge, S.; Zhou, S.K.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. https://doi.org/10.1016/j.media.2022.102510
100. Tanida, T.; Müller, P.; Kaissis, G.; Rueckert, D. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023. https://github.com/ttanida/rgrg
101. Hou, W.; Cheng, Y.; Xu, K.; Li, W.; Liu, J. Recap: Towards precise radiology report generation via dynamic disease progression reasoning. arXiv preprint **2023**, arXiv:2310.13864. https://github.com/wjhou/Recap
102. 102. Li, Y.; Yang, B.; Cheng, X.; Zhu, Z.; Li, H.; Zou, Y.; et al. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2023; pp. 2851–2861. https://doi.org/10.1109/ICCV51070.2023.00268
103. Nooralahzadeh, F.; Perez-Gonzalez, N.; Frauenfelder, T.; Fujimoto, K.; Krauthammer, M. Progressive transformer-based generation of radiology reports. arXiv preprint **2021**, arXiv:2102.09777. https://github.com/uzh-dqbm-cmi/ARGON
104. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2002; pp. 311–318.
105. Lin, C.Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (ACL)*; Barcelona, Spain, 2004.
106. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*; 2020. https://github.com/Tiiiger/bert_score
107. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2015; pp. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

 doi:10.20944/preprints202511.0010.v1

108. Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Pontes-Reis, E.; et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* **2023**, *4*, 100802. https://doi.org/10.1016/j.patter.2023.100802
109. Kaur, N.; Mittal, A.; Singh, G. Methods for automatic generation of radiological reports of chest radiographs: A comprehensive survey. *Multimed. Tools Appl.* **2022**, *81*, 13409–13439. https://doi.org/10.1007/s11042-021-11272-6
110. Mahmoud, M.; Monshi, A.; Poon, J.; Chung, V. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* **2020**, *106*, 101878. https://doi.org/10.1016/j.artmed.2020.101878
111. Liao, Y.; Liu, H.; Spasic, I. Deep learning approaches to automatic radiology report generation: A systematic review. *Informatics Med. Unlocked* **2023**, *39*, 101273. https://doi.org/10.1016/j.imu.2023.101273