Expert Systems With Applications 331 (2026) 133320

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Automated radiology report generation: A comprehensive review

Lina Huang *, Tasin Islam , Alina Miron , Kate Hone , Yongmin Li

Department of Computer Science, Brunel University London, Uxbridge, London, UB8 3PH, UK

ARTICLE INFO

Keywords:
ARRG
Radiology report formats
Foundation models
Multimodal
Knowledge integration
Ethical challenges

ABSTRACT

As the workload of radiologists continues to increase, writing radiology reports remains a time-consuming and error-prone task. In recent years, Automated Radiology Report Generation (ARRG) has emerged as a research hotspot aimed at addressing this challenge. This review analyses the main ARRG research streams, including template-based, retrieval-based, encoder-decoder, foundation-model-based, and hybrid approaches. We trace the evolution of ARRG from early template and retrieval paradigms to encoder-decoder and more recent foundation-model-based approaches, while also discussing the growing roles of multimodal, knowledge integration and reinforcement learning strategies, and we compare their respective strengths and limitations. We further summarise the commonly used public, restricted, and private datasets in ARRG research, while distinguishing between datasets that can directly support report generation and auxiliary resources mainly used for pretraining, grounding, or evaluation. In addition, we examine the clinical implications of radiology report format, with particular attention to the trade-offs between free-text and structured reporting and their consequences for model design. We also review mainstream evaluation methods for ARRG, including quantitative metrics (e.g., NLG and CE metrics) and qualitative assessment, and discuss why factual correctness, report organization, and clinical usefulness are not fully captured by surface-level language similarity alone. Finally, we discuss ethical and governance issues that are especially salient for ARRG, such as privacy, bias, hallucination, omission of key abnormalities, negation errors, and responsibility allocation in clinical workflows. We hope that this review will serve as a useful reference for future ARRG research and for the safe translation of these systems into clinical practice.

1. Introduction

Medical imaging, encompassing modalities such as radiography, CT, MRI, and ultrasound, is a noninvasive diagnostic tool that provides information about internal structures, tissues, and organs of the human body, including the location, morphology, and metabolic characteristics of lesions, without causing physical harm to the patient (Achenbach et al., 2022; Beddiar et al., 2023; Liao et al., 2023; Wang et al., 2024b). This method is highly repeatable, making it suitable for monitoring disease progression and changes over time. For instance, in tumour follow-ups or treatment evaluations, medical imaging can dynamically monitor disease progression and assess therapeutic efficacy by comparing pre- and post-treatment images (Aerts et al., 2014; Emaminejad et al., 2015; Popovici et al., 2017; Scalco & Rizzo, 2017). Therefore, medical imaging plays a critical role in clinical practice, providing indispensable support for accurate diagnosis, treatment guidance, as well as disease screening and prevention (Reale-Nosei et al., 2024). In diagnosis, medical imaging enables the clear identification and evaluation of abnormalities such as tumours, vascular diseases, and fractures, offering high-resolution images as a basis for clinical decisions. During treatment, imaging technology assists doctors in precisely locating lesions, guiding surgical procedures or radiation therapy plans, ensuring accurate radiation dose distribution in the target area, and maximizing the protection of healthy tissues (Lecchi et al., 2008). In disease screening and prevention, medical imaging is widely applied in early detection programs, such as mammography for breast cancer screening and low-dose CT for lung cancer screening, facilitating early diagnosis and intervention to reduce mortality rates (Javaid et al., 2024; Mazzone et al., 2018; Ren et al., 2022). Thus, with its efficiency, safety, and high repeatability, medical imaging holds a pivotal position in modern clinical medicine.

The essential role of medical imaging in clinical diagnosis and treatment has led to a continuous increase in the number of imaging examinations, posing significant challenges to healthcare systems. The interpretation of the massive volume of medical imaging data relies on professional radiologists. However, the global shortage of radiologists has become increasingly severe. For instance, according to statistics (The Royal College of Radiologists, 2023), in 2023, the UK

* Corresponding author.
E-mail addresses: lina.huang@brunel.ac.uk (L. Huang), tasin.islam@brunel.ac.uk (T. Islam), alina.miron@brunel.ac.uk (A. Miron), kate.hone@brunel.ac.uk (K. Hone), yongmin.li@brunel.ac.uk (Y. Li).

https://doi.org/10.1016/j.eswa.2026.133320
Received 7 June 2025; Received in revised form 16 April 2026; Accepted 13 June 2026
Available online 15 June 2026

clinical radiology workforce grew by 6.3%, but the demand for CT and MRI reporting surged by 11%. In the same year, the UK faced a 30% shortage of clinical radiologists, which is projected to rise to 40% by 2028 (The Royal College of Radiologists, 2023). The growing demand for imaging examinations has resulted in excessive workloads, reduced reading times, and fatigue among radiologists, sometimes preventing them from fulfilling their duties satisfactorily (Kaur et al., 2022). Some previous studies (Gatt et al., 2003; Marrie, 1997) have shown that delays in the generation of imaging reports have, in some cases, led to reports being interpreted by general physicians rather than trained radiologists. Discrepancies between interpretations by clinical physicians and radiologists can occasionally be significant, potentially affecting the treatment process for some patients (Benger & Lyburn, 2003; Kaur et al., 2022).

In response to this challenge, research on automated radiology report generation (ARRG) has garnered widespread attention. ARRG leverages deep learning technologies to automatically analyse medical images and generate reports, alleviating the workload of radiologists, accelerating the report generation process, and improving diagnostic efficiency. It can also serve as a guidance tool for less experienced radiologists, helping them identify potential abnormalities and reducing the risk of missed diagnoses (Liu et al., 2021a; Yang et al., 2022; You et al., 2021). Some people (Liao et al., 2023) suggest that the earliest work was in 2015, when Shin et al. utilized natural language processing (NLP) techniques to extract semantic labels from radiology reports and employed deep convolutional neural networks (CNNs) to map CT/ MRI images to different topic categories (Shin et al., 2016a). Additionally, they developed a model to generate keyword descriptions associated with CT/ MRI images and predict the presence of common diseases in the images. Other studies (Sloan et al., 2024), however, identify 2016 as the starting point, coinciding with the release of the Indiana University X-ray Dataset (IU X-ray) (Demner-Fushman et al., 2016) and Shin et al. utilized this dataset to develop a method for automatically annotating medical images by employing a CNN as the encoder and a recurrent neural network (RNN) as the decoder (Shin et al., 2016b). Subsequently, by 2018, research on ARRG began to gain widespread attention. From 2020 onwards, the number of publications in this field has grown rapidly, and it is expected that this trend will become increasingly prominent in the future (Sloan et al., 2024).

ARRG, as a specialized branch of image captioning (Liao et al., 2023; Luan et al., 2023; Sloan et al., 2024; Wang et al., 2022a), focuses on generating diagnostic reports for medical images, such as X-rays, CT scans, MRIs, and ultrasounds. It achieves this by leveraging the combined strengths of computer vision (CV) (Elyan et al., 2022; Javaid et al., 2024) and natural language processing (NLP) (Chng et al., 2023; Zhang et al., 2019b).

Before beginning ARRG research, it is crucial to determine the format of the diagnostic report to be generated, as this will influence the design of the ARRG model. For instance, if the goal is to produce a structured report, the model design might incorporate template-based methods or draw inspiration from template-based approaches, possibly adopting hybrid methods.

As for the data, ideal ARRG research requires high-quality medical images, which involves two key aspects. On the one hand, the medical images must be clear and meet the technical standards for radiological examinations. For example, a qualified chest X-ray (Chand et al., 2013) should fully display both lung apices and costophrenic angles, ensuring that the distances between the inner edges of the clavicles and the spine are equal to prevent image rotation. On the other hand, an ideal dataset should also include complete and standardized diagnostic reports. These reports must be detailed and accurate, reflecting the pathological findings in the images and conforming to the standards of medical documentation. However, datasets that meet these criteria are exceedingly rare in practice.

From a methodological perspective, ARRG research can be understood along two complementary dimensions: methodological paradigms and enhancement mechanisms. The methodological paradigms include template-based (Abela et al., 2022; Pino et al., 2021), retrieval-based (Charalampakos et al., 2021; Kougia et al., 2021; Liao et al., 2023; Ni et al., 2020; Reale-Nosei et al., 2024; Syeda-Mahmood et al., 2020; Yang et al., 2021c; Zhang et al., 2018), encoder-decoder, foundation-model-based, and hybrid approaches. Among them, the encoder-decoder framework remains one of the most classical and influential development lines in ARRG (Liang et al., 2018; Shin et al., 2016b; Sloan et al., 2024; Vinyals et al., 2015; Yao et al., 2017). Early encoder-decoder systems (Shin et al., 2016b) often employed CNN-RNN architectures, typically combining CNN visual encoders with RNN, LSTM (Hochreiter, 1997), or GRU (Cho et al., 2014; Chung et al., 2014, 2015) decoders, and were later extended to CNN-Hierarchical RNN (or Hierarchical LSTM) variants (Harzig et al., 2019; Huang et al., 2019; Jing et al., 2017; Yin et al., 2019; Yuan et al., 2019; Zhang et al., 2020). Attention mechanisms were subsequently incorporated to improve image-text alignment (Gajbhiye et al., 2022; Jing et al., 2017; Song et al., 2022; Wang et al., 2018, 2022b; Yan et al., 2022; You et al., 2021). With the widespread success of Transformer (Vaswani, 2017) in image captioning (Herdade et al., 2019; Yu et al., 2019), Transformer-based and foundation-model-based ARRG approaches have become increasingly prominent (Alfarghaly et al., 2021; Chen et al., 2020; Lovelace & Mortazavi, 2020; Nooralahzadeh et al., 2021; Wang et al., 2023b; Zhou et al., 2022). These more recent approaches can be viewed as extending or reconfiguring earlier encoder-decoder ideas while also introducing new large-scale pretraining and cross-modal modeling capabilities. In contrast, attention, multimodal fusion, and knowledge integration are better understood as cross-cutting modeling strategies, while reinforcement learning is more appropriately regarded as a cross-cutting optimization strategy, because these elements may be incorporated into different model families in different ways rather than constituting fully separate report-generation paradigms.

Within this evolving methodological landscape, several trends are especially notable. One of the major challenges in ARRG research is the scarcity of medical data, which has contributed to the growing importance of foundation models (Bannur et al., 2023a; Boecking et al., 2022; Delbrouck et al., 2022; Li et al., 2023c, 2024b; Nicolson et al., 2023; Wu et al., 2023a). Pre-trained on large-scale datasets, these models can be adapted to relatively smaller domain-specific datasets, thereby partially alleviating the problem of limited medical data (Liao et al., 2023) while also supporting stronger cross-modal representation learning. In parallel, hybrid methods have emerged as an important direction because template-based, retrieval-based, and conventional encoder-decoder methods each have limitations in ARRG tasks (Biswal et al., 2020; Gao et al., 2024; Nie & Liu, 2023; Tanwani et al., 2022; Yang et al., 2021c). By combining the strengths of different paradigms, such methods can improve both the reliability and the expressive flexibility of generated reports.

Another important direction in ARRG research is the integration of medical knowledge and multimodal information. Some studies incorporate medical knowledge through techniques such as knowledge distillation (Huang et al., 2023; Liu et al., 2021a), knowledge embedding (Liu et al., 2021b), or knowledge graphs (Li et al., 2024b; Wang et al., 2023a; Zhang et al., 2020), with the aim of improving factual grounding and report accuracy (Gajbhiye et al., 2022; Hou et al., 2023b; Huang et al., 2023; Kim et al., 2023; Li et al., 2022; Liu et al., 2021a,b; Nishino et al., 2022). Meanwhile, because radiologists often rely on patients' medical history, previous examinations, and relevant laboratory results when writing diagnostic reports, these real-world workflows have motivated the incorporation of multimodal information into ARRG systems (Dalla Serra et al., 2022; Jeong et al., 2024; Liu et al., 2021a; Mondal et al., 2023; Nguyen et al., 2023, 2021; Shang et al., 2022).

Most ARRG studies have focused on 2D imaging. However, compared with 2D imaging (e.g., X-rays), 3D imaging (e.g., CT and MRI) can provide a more comprehensive representation of a patient's condition. As a result, some studies have explored ARRG methods for 3D medical

Table 1
Comparison of representative prior reviews related to ARRG and the positioning of the present review.

Review	Year	Main emphasis	Report-format implications	Direct vs. auxiliary dataset roles	Method organization beyond simple taxonomy	Task-specific ethics synthesis	Position relative to the present review
(Monshi et al., 2020)	2020	Early survey of automatic radiology report generation methods, datasets, and evaluation	Limited	No	Mainly architecture-oriented	No	Valuable early overview, but covered research stages are too early, less emphasis on data stratification, task-specific ethical risks
(Kaur et al., 2022)	2022	Deep-learning research on radiology report generation, especially chest X-ray settings	Partial	No	Mainly method-focused	No	Stronger focus on radiology reporting than broader medical-report surveys, but less emphasis on data stratification, task-specific ethical risks
(Messina et al., 2022)	2022	Explainability, evaluation, and physician-centred assessment for automatic report generation	Partial	Partial	Partial	Limited	Provides important evaluation and explainability perspectives, but less emphasis on task-specific ethics synthesis
(Liao et al., 2023)	2023	Broad survey of deep learning approaches, datasets, and challenges in ARRG	Partial	Yes	Yes	Limited	Covers many ARRG components, but less emphasis on task-specific ethics synthesis
(Liu et al., 2023a)	2023	Recent advances in datasets, models, and evaluation	Limited	Limited	Partial	No	Useful update-oriented review, but without task-specific ethical risks
(Pang et al., 2023)	2023	A review of traditional ARRG frameworks	No	No	Mainly method-focused	No	Intuitive ARRG structure but no data stratification and task-specific ethics synthesis
(Sloan et al., 2024)	2024	Recent ARRG developments, including multimodal methods and evaluation	Partial	No	Yes	Partial	More current than earlier surveys, but still less focus on data stratification and task-specific ethical risks
(Liu et al., 2025)	2025	State-of-the-art medical report generation beyond radiology-specific settings	Partial	Limited	Broad taxonomy	Limited	Provides wider generative-medical-report context, but less emphasis on task-specific ethics synthesis
Present review	**2025**	ARRG-focused synthesis spanning report formats, datasets, methods, evaluation, and ethics	**Yes**	**Yes**	**Paradigms + mechanisms**	**Yes**	**Emphasizes report-format implications, dataset-role stratification, comparative methodological synthesis, and ARRG-specific generative ethics risks**

imaging, such as RadFM (Wu et al., 2023b), CT2Rep (Hamamci et al., 2024c), and M3D-LaMed (Bai et al., 2024). Nevertheless, ARRG for 3D imaging still faces substantial challenges, including data processing complexity, slice aggregation, and computational cost.

Evaluation is also a critical aspect of ARRG research. Common evaluation methods currently include quantitative metrics (such as natural language generation evaluation metrics and clinical efficacy evaluation metrics) and qualitative assessments. However, both types of evaluation methods have limitations, which will be discussed in detail in Section 6.

As ARRG technology advances, ethical and safety issues are becoming increasingly important, including data privacy, algorithmic bias, clinical responsibility, and interpretability. These issues warrant sufficient attention. For instance, strict data privacy protection measures, such as data anonymization and access controls, are needed to ensure patient privacy. Training data should be analysed and processed to minimize bias. The auxiliary role of ARRG models should be clearly defined to prevent over-reliance, and corresponding clinical responsibility mechanisms should be established.

Existing reviews have already surveyed ARRG or closely related medical-report-generation research from different perspectives. One line of review work has focused on radiology report generation itself, discussing deep-learning methods, datasets, and evaluation (Monshi et al., 2020; Pang et al., 2023). Some reviews were more specifically centered on chest radiographs or chest X-ray report generation (Kaur et al., 2022). Some broader surveys covered automatic medical imaging or medical report generation at a wider scope beyond radiology-specific report generation (Liao et al., 2023; Liu et al., 2025). Other reviews placed particular emphasis on explainability, evaluation, and physician-centred assessment (Messina et al., 2022). More recent reviews have further summarized recent advances in ARRG, including developments in datasets, model design, multimodal methods, and evaluation (Liu et al., 2023a, 2025; Sloan et al., 2024).

However, the emphasis of these reviews differs. Some focus primarily on methodological development and model architectures (Kaur et al., 2022; Liao et al., 2023; Liu et al., 2023a; Monshi et al., 2020; Sloan et al., 2024), some place greater emphasis on explainability and evaluation (Messina et al., 2022), and some discuss medical report generation at a broader level without focusing specifically on report-format implications for ARRG system design (Liu et al., 2023a). In addition, although prior reviews discuss datasets and evaluation (Kaur et al., 2022; Pang et al., 2023; Sloan et al., 2024), they do not always organize these resources in a way that explicitly distinguishes direct report-generation datasets from auxiliary resources such as label-only, grounding, or graph-annotated datasets. Issues related to safety, interpretability, and data-related concerns are also occasionally acknowledged in prior reviews (Kaur et al., 2022; Liu et al., 2023a; Monshi et al., 2020; Pang et al., 2023), but are usually not developed in a task-specific way for generative radiology reporting. Table 1 summarizes these differences more explicitly. Against this background, our review aims to make the following contributions:

1. *Provides a more task-oriented review of data resources for ARRG*

   Compared with prior reviews that discuss datasets more broadly or without consistently separating their functional roles (Kaur et al., 2022; Monshi et al., 2020; Pang et al., 2023; Sloan et al., 2024), this review focuses specifically on radiology-oriented image-text resources for ARRG, including X-ray, CT, MRI, ultrasound (in multimodal settings), and PET/CT. More importantly, we distinguish between datasets that can be used directly for report generation and auxiliary resources that are mainly useful for pretraining, weak supervision, anatomical grounding, knowledge-graph construction, or evaluation. We also clarify why pathology data, although clinically important as a diagnostic gold standard, usually plays a different role from radiology report data in the ARRG pipeline.
2. *Organizes ARRG methods with greater emphasis on architectural evolution and enhancement mechanisms*

   Compared with prior reviews that are mainly architecture-oriented or present methods largely category by category (Kaur et al., 2022; Monshi et al., 2020; Pang et al., 2023), we discuss ARRG methods along two complementary dimensions: methodological paradigms and enhancement mechanisms. This structure highlights the historical encoder-decoder thread while still accommodating template-based, retrieval-based, foundation-model-based, and hybrid paradigms. We further discuss attention, multimodal fusion, knowledge integration, and reinforcement learning as mechanisms that may act either as central architectural components or as add-ons, depending on the model family. Tables 7, 9, and 10 are used not simply to list scores but to support discussion of how model design choices interact with evaluation protocols.
3. *Delivers a clinically oriented evaluation of ARRG applications, highlighting the impact of radiology report formats on system design and arguing that hybrid reporting targets may be a promising direction*

   Compared with broader medical-report surveys and method-focused ARRG reviews that do not place primary emphasis on report-format implications (Liu et al., 2023a; Pang et al., 2023), this review provides an in-depth exploration of the clinical value of ARRG from a medical perspective. Through comparison, we evaluate the application performance of free-text radiology format versus structured radiology format across different clinical scenarios, while assessing their respective advantages and limitations. The radiology report format significantly influences ARRG architectural design. For instance, structured reports, with their standardized nature, are particularly well-suited for template-based approaches. This architecture ensures completeness and consistency of key clinical elements, as demonstrated in standardized reporting systems like the Breast Imaging Reporting and Data System (BI-RADS). Current clinical evidence confirms that both free-text and structured formats possess irreplaceable value. Accordingly, we argue that future ARRG systems may benefit from hybrid reporting targets that combine natural-language flexibility for complex or rare pathologies with structured anchors for critical diagnostic elements (e.g., lesion location, size, and morphological characteristics). Such adaptive systems could better align with clinical needs while accommodating the diversity of disease manifestations.
4. *Expands the ethical discussion by connecting general medical AI concerns to ARRG-specific generative risks*

   Prior reviews have rightly highlighted safety, explainability, and clinical responsibility in medical AI (Liao et al., 2023; Liu et al., 2025; Messina et al., 2022; Sloan et al., 2024). Building on that broader discussion, this review devotes dedicated attention to ethical questions that are especially salient in ARRG, including hallucinated findings, failure to preserve negation, omission of clinically important abnormalities, inconsistency between Findings and Impression, misuse of prior-study references, and the way apparently fluent text may obscure clinically unsafe content. Rather than claiming absolute novelty, we position this discussion as a more explicit and task-specific synthesis of ethical issues for generative radiology reporting.

Fig. 2 outlines key aspects of ARRG research, including review methodology, radiology report formats, experimental datasets, ARRG methodological paradigms and enhancement mechanisms, model evaluation methods, and ethical issues. The remaining structure of this article is as follows: Section 2 describes the review methodology adopted in this work, including the literature scope, search sources, time window, and inclusion and exclusion criteria. Section 3 provides a detailed introduction to the two commonly used formats of radiology reports, discussing their respective advantages, disadvantages, and future trends. Section 4 reviews the datasets currently used in ARRG research, including X-ray, CT, MRI, ultrasound, and PET/CT radiological imaging data, while analysing the characteristics of different data sources and their applications in ARRG tasks. Section 5 discusses ARRG through two complementary lenses: methodological paradigms and enhancement mechanisms. Under this structure, template-based, retrieval-based, encoder-decoder, foundation-model-based, and hybrid approaches are treated as methodological paradigms, while attention mechanisms, multimodal fusion, knowledge integration, and reinforcement learning are discussed as mechanisms that may operate across different model families. Section 6 explores the evaluation methods for ARRG, including quantitative evaluation metrics (such as natural language generation (NLG) evaluation metrics, clinical efficacy (CE) evaluation metrics, etc) and qualitative evaluation methods (such as human evaluation, etc) to assess the quality and usability of generated reports. Additionally, this section compares the experimental results of existing ARRG studies to identify the state-of-the-art methods. Section 7 independently discusses ethical issues in ARRG research, focusing on key challenges such as data privacy protection, model fairness, responsibility allocation, patient consent, and task-specific generative risks. Section 8 provides a discussion of the limitations of current ARRG research and presents an outlook on future trends in ARRG research. Section 9 summarizes the key contributions of this review.

2. Review methodology

This article is intended as a comprehensive and transparent narrative review rather than a formal systematic review or meta-analysis. To improve transparency and reproducibility, the literature identification process was guided by four principles: broad coverage of relevant sources, explicit topic keywords, relevance to radiology report generation, and backward/forward reference checking from influential papers and prior reviews. The main database search covered PubMed, Scopus, IEEE Xplore, Web of Science, and Google Scholar was used as an additional source to identify potentially relevant records. The review window covered studies published from January 1, 2015 to April 14, 2026, which corresponds to the period in which ARRG emerged as a recognizable research topic and then developed rapidly. The search scope focused on peer-reviewed articles and widely used preprints related to automated radiology report generation. Representative search terms included combinations of “automated radiology report generation”, “radiology report generation”, “medical report generation”, “structured reporting”, “image-to-text”, “multimodal radiology”, “knowledge graph”, and “large language model”, together with modality-specific terms such as “chest X-ray”, “CT”, and “MRI”.

The screening process was performed iteratively. In total, 804 records were identified from the main databases and 94 additional records were identified through Google Scholar, yielding 898 records overall. After removing 279 duplicates, 619 records remained for title and abstract screening. At this stage, 431 records were excluded. The remaining 188 full-text reports were assessed for eligibility, and 64 were excluded with reasons. Ultimately, 124 studies were included in the core ARRG analytical corpus for methodological and experimental synthesis. This number refers specifically to ARRG-focused studies analysed in depth, rather than to the total number of references cited in this review. Backward and forward citation tracing from influential ARRG studies,

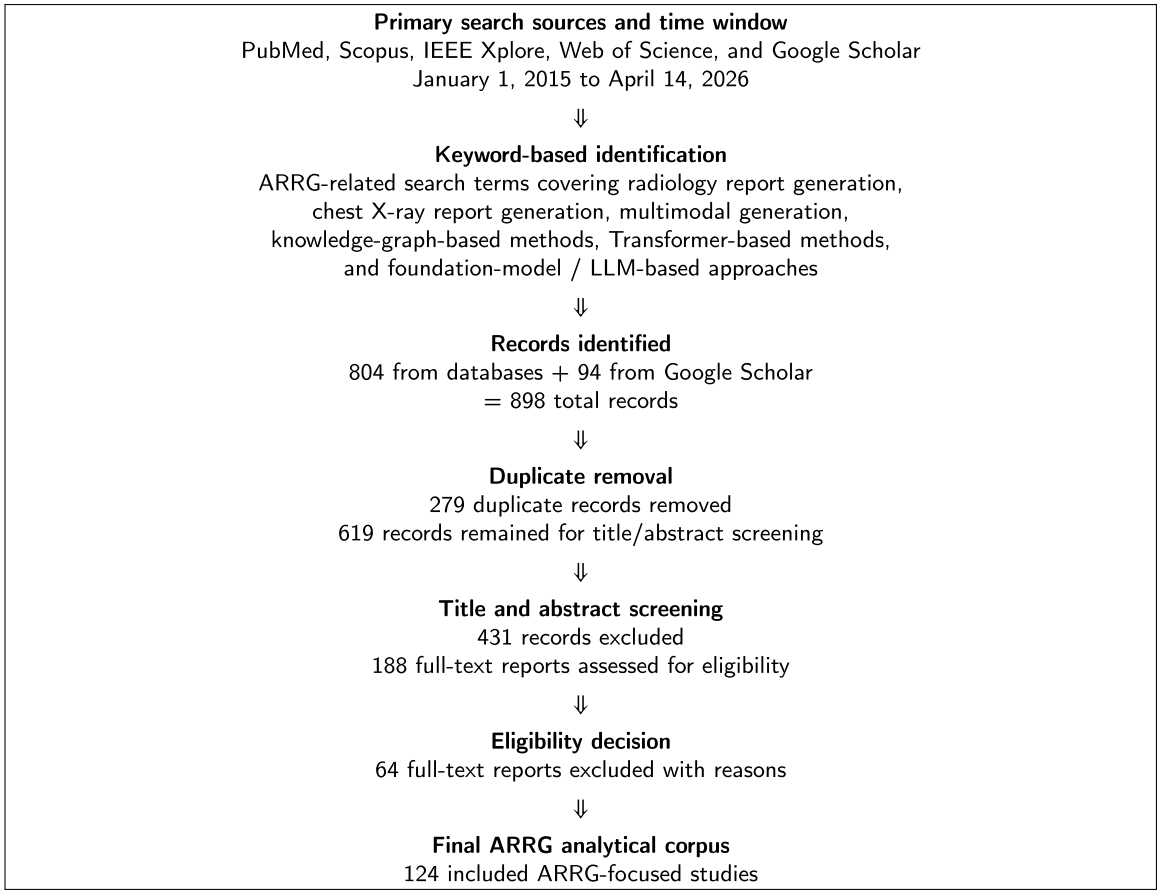


Fig. 1. Schematic summary of the literature identification and screening process used in this review.

dataset papers, and prior reviews was also used to improve coverage and reduce the risk of missing influential work.

Inclusion criteria

1. Studies focused directly on automated radiology report generation or closely related radiology-specific report generation tasks.
2. Studies making a methodological, data-related, evaluative, or clinically relevant contribution that directly informed the analysis of ARRG systems, datasets, report organization, or assessment.
3. Publications containing sufficient technical or clinical detail to support analysis.

Exclusion criteria

1. Studies focused solely on non-radiology domains such as pathology or ophthalmology or metrics evaluation without a clear radiology-report-generation connection.
2. Purely image-classification or segmentation studies with no direct ARRG relevance, except where such datasets were subsequently reused as auxiliary resources.
3. Duplicated records across search sources or articles lacking sufficient information for reliable interpretation.

Fig. 1 provides a schematic summary of the literature identification and screening logic used in this review. The final corpus was organized into five analytical themes: report formats, datasets, methods, evaluation, and ethics. Within the dataset section, we explicitly distinguish between direct report-generation datasets and auxiliary resources. Within the methods section, we distinguish between methodological paradigms and enhancement mechanisms. This design choice reflects the fact that ARRG is a rapidly evolving field in which many techniques, such as attention or multimodal fusion, function differently across model families. We also note that the field is moving quickly; therefore, despite efforts to keep the review up to date, newly released datasets and models may emerge after the review window covered in this manuscript.

3. Report formats

Radiology reports serve as a critical communication tool between radiologists and referring physicians (Bécares-Martínez et al., 2020; Caranci et al., 2020; European Society of Radiology (ESR), 2011; Ierardi et al., 2020; Neri et al., 2020; Segrelles et al., 2017; Sobez et al., 2019). An effective report should accurately address clinical questions, use clear, precise, and unambiguous language, provide explicit management recommendations, and include relevant negative findings when appropriate (Bécares-Martínez et al., 2020; Caranci et al., 2020; European Society of Radiology (ESR), 2011, 2023; Ierardi et al., 2020; The Royal College of Radiologists, 2023).

As mentioned in the introduction, before conducting ARRG research, it is essential to define the expected report format. The report format can influence the design of the ARRG system and generation strategy to some extent. However, there is currently no globally unified format for radiology reports. The most common formats can be categorized into traditional free-text reporting and structured reporting (European Society of Radiology (ESR), 2011, 2018, 2023; Neri et al., 2022).

Free-text reporting refers to diagnostic reports written in a narrative, prose-style, unstructured format (European Society of Radiology (ESR), 2011, 2018, 2023; Neri et al., 2022). In contrast, structured reporting is widely regarded as a potential method for improving the quality of radiology reports (Dick et al., 2021; Ganeshan et al., 2018; Reiner, 2009), with its benefits extensively discussed and recognized (European Society of Radiology (ESR), 2023; Powell & Silberzweig, 2015). In 2018, the European Society of Radiology (ESR) identified three key reasons for

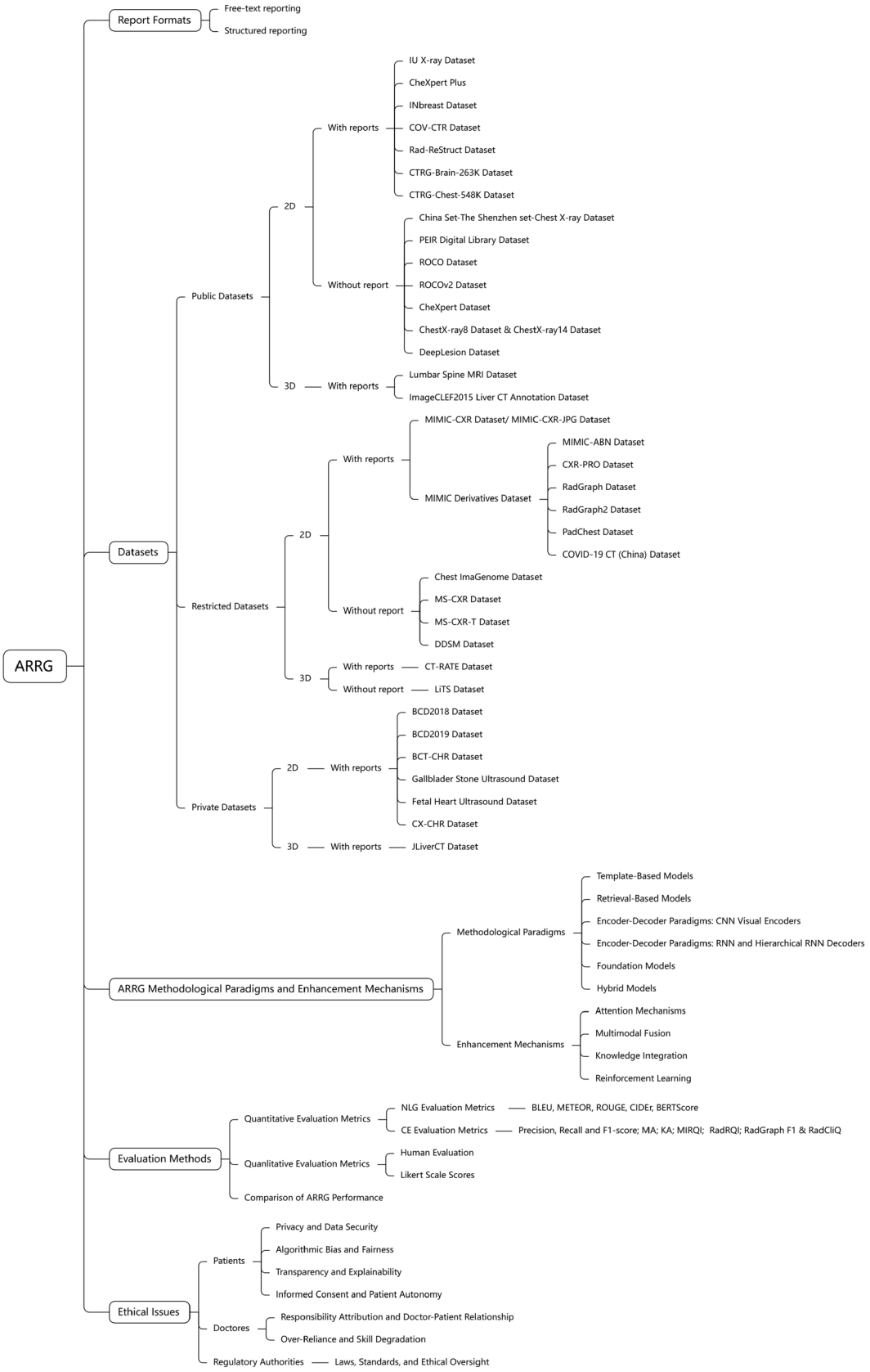


Fig. 2. An outline of key aspects related to ARRG research.

transitioning from traditional free-text reporting to structured reporting: quality, datafication/quantification, and accessibility. Both the ESR and the Radiological Society of North America (RSNA) have strongly promoted the adoption of structured reporting (European Society of Radiology (ESR), 2018, 2023; Morgan et al., 2014). The RSNA developed the RadReport Template Library (RSNA, 2025) as part of its Structured Reporting Initiative. However, it is no longer being updated.

It is important to note that the term "structured reporting" has been somewhat ambiguous. In the past, it was often used interchangeably with "standardized reporting." In 2020, Nobel et al. proposed the following definitions: structured reporting refers to the organization of radiology report content in a structured manner through IT-based means, while standardized reporting refers to the optimization and unification of the content in radiology reports (Nobel et al., 2020). Before this, no clear distinction was made between standardized reporting and structured reporting, and the two terms were often considered synonyms. Nobel et al. argue that true structured reporting involves the use of specialized IT solutions to store report content in a structured format, facilitating data mining (Nobel et al., 2020). In contrast, standardized reporting emphasizes the high degree of content and language standardization, as well as the structured organization of specific sections of a report, without necessarily relying on IT tools. Additionally, Nobel et al. divided structured reporting into two levels (Nobel et al., 2020): Level 1 refers to structured layout, which presents results in a strict, predefined order, creating and maintaining consistency, similar to a template or blueprint for a report. Level 2 refers to structured content, which involves the organization and presentation of medical content within the report and represents the more technical aspect of IT-guided content generation. Examples mentioned in the literature are dropdown menus, pick lists, and point-and-click systems. Nobel et al. emphasized that establishing standardized structures should precede the implementation of structured reporting to maximize clinical practice benefits (Nobel et al., 2020). They also noted that "Standardized reporting is about report content, structured reporting is an IT-based tool".

In the ESR's 2018 publication on structured reporting in radiology (European Society of Radiology (ESR), 2018), based on the description by (Weiss & Bolos, 2010), structured reporting was divided into three levels: structured format, including paragraphs and subheadings; organizational consistency; and the use of professional terminology, i.e., standardized language. In the ESR's 2023 update (European Society of Radiology (ESR), 2023), it was further stated that: "In contrast, a radiological report can be highly standardized with regard to content and language (standardized reporting) or structured in specific sections but without the use of a supporting IT tool (structured reporting Level 1 (Nobel et al., 2020))." Thus, whether following (Nobel et al., 2020), who argue that standardized structures should precede structured reporting, or the ESR's 2018 (European Society of Radiology (ESR), 2018) and 2023 (European Society of Radiology (ESR), 2023) publications on structured reporting in radiology, the consensus is clear: the content of radiology reports should be standardized. In fact, much of the research on structured reporting in radiology focuses on report content, without strictly distinguishing between standardized reporting and structured reporting (with IT tools). For example, some studies regard BI-RADS (Breast Imaging-Reporting and Data System) (D'Orsi et al., 2013) as a classic example of structured reporting (Ganeshan et al., 2018).

Given that the focus of ARRG research is on the content of the generated report, rather than whether IT tools are used for importing, in this literature review, the concept of structured reporting is broader. It includes not only Level 1 and Level 2 structured reporting, but also standardized reporting, such as reports based on the classic BI-RADS (D'Orsi et al., 2013) structured template, which are also considered structured reporting (generalized structured reporting) in this literature review because they all use standardized terminology and formats.

For the purposes of ARRG, this broader operational definition is more useful than a purely terminological debate. The central modeling question is whether the target report exhibits explicit section structure, standardized terminology, constrained attribute slots, or freer narrative expression. In other words, the issue is not only how the report is entered in the hospital information system, but how its linguistic and semantic structure shapes supervision, decoding, and downstream evaluation. More standardized and structurally constrained targets are generally easier to supervise and evaluate automatically, whereas freer report styles preserve descriptive flexibility but make alignment, generation control, and evaluation more difficult.

The reason why there is currently no unified radiology report format worldwide is that both free-text reporting and structured reporting formats have their own advantages and disadvantages, and neither is perfect. As a crucial department in medical technology, the goal of the radiology department is to provide clinically valuable imaging information and professional interpretation to clinicians, ultimately collaborating with clinical departments to create the best diagnostic and treatment plan for patients. Clinicians, in turn, combine the patient's medical history, physical examination, and other information to make an overall judgment on the condition and reach a final diagnosis and treatment decision.

The ESR suggests that an ideal radiology report should include the following sections (European Society of Radiology (ESR), 2011):

1. Clinical Referral: A brief overview of the indication for the referral and the underlying reasons. This section typically aims to answer the clinical question posed by the referring physician.
2. Technique: A detailed description of the type of examination performed and the imaging techniques or sequences used.
3. Findings: A comprehensive description of all imaging observations, prioritizing those most relevant to the referral indication, but also including incidental findings. In some cases, negative findings are also highlighted to help rule out certain differential diagnoses.
4. Impression/ Conclusion: A concise summary of the observations, usually consisting of one or two sentences. This often includes a prioritized list of diagnoses or differential diagnoses to guide subsequent patient management.
5. Advice: The report may provide recommendations for further action. For example, the report may suggest further investigations to help clarify the diagnosis.

To understand the advantages and disadvantages of free-text reporting and structured reporting, let us consider the example of a mammography report. For the structured report of a mammogram, the classic BI-RADS (D'Orsi et al., 2013) template will be used. In fact, some radiologists strictly follow the BI-RADS (D'Orsi et al., 2013) structured template when writing mammography reports, such as those in the Radiology Department of Xiamen University First Affiliated Hospital China, which established a Breast Imaging Diagnostic Center. The report template used by this department is shown in Table 2 (the original template is in Chinese, while Table 2 displays the translated English version).

To better understand BI-RADS-based structured reports, let us briefly explain BI-RADS (D'Orsi et al., 2013). BI-RADS stands for the Breast Imaging-Reporting and Data System, a standardized reporting system developed by the American College of Radiology (ACR) aimed at improving the quality and consistency of breast imaging reports. BI-RADS also provides a standardized format for breast imaging reports, which includes the following sections:

- Examination name
- Clinical history and indications
- Breast composition
- Findings (including masses, calcifications, architectural distortion, asymmetries, intramammary lymph nodes, skin lesions, solitary dilated ducts, and associated features)
- Impression (categorization of lesions based on imaging features and assessment using the BI-RADS (D'Orsi et al., 2013) classification system (categories 0–6))

Table 2
Breast imaging structured report template (based on BI-RADS D'Orsi et al., 2013).

Department of Diagnostic Radiology		
Name: **Age:** **Sex:** **Date:** **Patient ID:** **Referring doctor:**		
BILATERAL/UNILATERAL (RIGHT/LEFT) BREAST MAMMOGRAPHY		
Mammography of both/right/left breast was performed using low dose radiation technique with adequate compression. Bilateral breast/Right breast/Left breast craniocaudal and mediolateral oblique views have been obtained.		
Additional mammography views were obtained (if any) (specify the view and the laterality).		
Previous mammogram dated (specify), reported as (BI-RADS 0/1/2/3/4/5/6) is available for comparison.		
Clinical presentation: **Breast composition:** **Finding:**		ACR Category a/ ACR Category b/ ACR Category c/ ACR Category d
Location of lesion:		
Mass:	Number: Size (cm): Shape: Margins: Density:	None/ Oval/ Round/ Irregular None/ Circumscribed/ Obscured/ Microlobulated/ Indistinct/ Spiculated None/ High/ Equal/ Low/ Fat-containing
Calcifications:	Typically benign: Suspicious: Distribution:	None/ Skin/ Vascular/ Coarse or "popcorn-like"/ Large rod-like/ Round/ Rim/ Dystrophic/ Milk of calcium/ Suture None/ Amorphous/ Coarse heterogeneous/ Fine pleomorphic/ Fine linear or fine-linear branching None/ Diffuse/ Regional/ Grouped/ Linear/ Segmental
Architectural distortion:		None/ Present
Asymmetries:		None/ Global/ Focal/ Developing
Intramammary lymph node:		None/ Present
Skin lesion:		None/ Present
Solitary dilated duct:		None/ Present
Associated features:		Skin retraction/Nipple retraction /Skin thickening/Trabecular thickening /Axillary adenopathy/ Architectural distortion/ Calcifications
Impression:		
BI-RADS:		0/ 1/ 2/ 3/ 4A/ 4B/ 4C/ 5/ 6
Advice:		Recall for additional imaging and/or comparison with prior examination(s)/ Routine mammography screening/ Routine mammography screening/ Short-interval (6-month) follow-up or continued surveillance mammography/ Tissue diagnosis/ Surgical excision when clinically appropriate

- Advice (recommendations based on the BI-RADS (D'Orsi et al., 2013) assessment, such as short-term follow-up, biopsy, etc.)

It is clear that the BI-RADS-based report template fully aligns with the European Society of Radiology's requirements for an ideal radiology report. To visually compare free-text reporting and structured reporting, let us now look at how the same mammography results would be written in both formats.

Case: *Female, 45 years old. Six months ago, the patient noticed non-bloody discharge from the left breast. Initial mammography revealed heterogeneous calcifications in the anterior region of the upper outer quadrant at the 1 o'clock position of the left breast. The findings were categorized as BI-RADS® Category 3, indicating a probably benign lesion, with a recommendation for follow-up in 6 months. The patient is now undergoing a scheduled follow-up mammogram.*

Tables 3 and 4 present the mammography results of the patient in free-text format and structured format, respectively.

By utilizing two different reporting formats for the same case (mammography examination), we can directly compare their differences. It is evident that structured reporting adopts a "checklist" approach, ensuring all questions posed by the referring clinician are comprehensively addressed using concise language. In contrast, free-text reporting is often hindered by verbose or ambiguous descriptions, requiring readers to review the entire report to extract treatment-related information. Moreover, structured reporting employs standardized terminology, which effectively avoids ambiguity (European Society of Radiology (ESR), 2018; Khurana et al., 2020) and facilitates comparative studies on disease treatment (European Society of Radiology (ESR), 2018). This provides clinicians with clearer guidance, thereby promoting appropriate care decisions (European Society of Radiology (ESR), 2018). Additionally, structured reporting enables easier comparisons in research and clinical practice (Dimarco et al., 2020; European Society of Radiology (ESR), 2018).

To further evaluate free-text and structured reporting, we examine findings from several studies. A study on the impact of the structured LI-RADS template for hepatocellular carcinoma (HCC) based on CT and MRI (Flusberg et al., 2017) found that structured reporting improved the completeness and consistency of key HCC features, reducing communication errors. Similarly, a study on whole-body CT scans (Schwartz et al., 2011) revealed that both clinicians and radiologists preferred structured reporting for its clarity and practicality, with higher satisfaction compared to free-text reports.

In the context of pancreatic ductal adenocarcinoma (PDA), structured reporting significantly enhanced inter-reader consistency in evaluating vascular involvement (e.g., portal vein, superior mesenteric artery) compared to free-text reporting (Dimarco et al., 2020). Free-text reports were more prone to ambiguity and omissions, potentially leading to misinterpretations of tumour resectability and unnecessary surgeries (Tem-

Table 3
Mammography report in free-text format.

UNILATERAL (LEFT) BREAST MAMMOGRAPHY
Mammography of left breast was performed using low dose radiation technique with adequate compression. Left breast craniocaudal and mediolateral oblique views have been obtained.
Compared with the previous mammogram dated 15/01/2023, reported as BI-RADS 3, there is an increase in the number of calcifications from the prior exam.
Clinical presentation: Non-bloody discharge left breast. **Finding:**
The left breast exhibits a fatty composition. No well-defined mass is observed within the left breast. Regionally distributed heterogeneous calcifications are noted in the upper outer quadrant of the left breast. The blood vessels of the left breast are symmetrical, and no abnormalities are detected in the pectoralis major muscle. The skin and areola of the left breast show no thickening, and there is no nipple retraction. The subcutaneous tissue structures are clearly visible. No significantly enlarged lymph nodes are observed in the left axilla.
Impression:
BI-RADS® Category 3: Probably Benign Finding. Short interval follow-up of the left breast is recommended in 6 months.

pero et al., 2021). Structured reporting also improved the practicality of staging reports, reducing the need for direct image review (Marcal et al., 2015). Additionally, a study on multiphase CT for pancreatic cancer (Brook et al., 2015) highlighted that structured reports provided more comprehensive surgical planning details, such as aberrant hepatic arteries and anatomical variations, which were often omitted in free-text reports.

Beyond pancreatic cancer, structured MRI reports for suspected rectal cancer (Nörenberg et al., 2017) improved surgical planning accuracy and increased referring surgeons' confidence in clinical decisions. Similarly, structured CT reports for colon cancer staging (Granata et al., 2022) were emphasized as essential for providing high-quality service to clinicians and patients while supporting research efforts. Overall, structured reporting has been shown to enhance patient outcomes through better surgical planning, earlier disease staging, and more comprehensive treatment (Rocha et al., 2020).

Overall, referring physicians generally prefer structured reporting over traditional free-text reporting (Schwartz et al., 2011; Sistrom & Honeyman-Buck, 2005), largely due to its higher clarity (Schwartz et al., 2011). Specifically, free-text reporting often varies in language, length, and descriptive style, making it difficult for clinicians to extract key patient care information (Ganeshan et al., 2018). Additionally, free-text reports may lack critical preoperative tumour staging information, leading to inappropriate patient management (Marcal et al., 2015). Structured reporting, on the other hand, offers several advantages. It ensures the completeness and comparability of reports (European Society of Radiology (ESR), 2018). It facilitates automated functions (e.g., TNM staging), integration with other clinical parameters (e.g., laboratory results), and data sharing (e.g., registries, biobanks) (European Society of Radiology (ESR), 2018). Structured reporting reduces ambiguity, improves communication with referring clinicians, and decreases misdiagnoses (European Society of Radiology (ESR), 2018, 2023; Khurana et al., 2020; Park et al., 2010; The Royal College of Radiologists, 2023). Moreover, structured reporting enhances the consistency (Dimarco et al., 2020) and reproducibility (Brook et al., 2015; Eghtedari et al., 2021; Flusberg et al., 2017; Neri et al., 2022; Nörenberg et al., 2017; Sahni et al., 2015) of radiology reports, improving their readability and clarity. This facilitates data extraction and mining for research and educational purposes (European Society of Radiology (ESR), 2018, 2023; Goel et al., 2019; Khurana et al., 2020; Nobel et al., 2020; Pinto dos Santos et al., 2018; The Royal College of Radiologists, 2023). Finally, structured report data greatly advances artificial intelligence (AI) development, as structured data are more readily used as training and validation datasets (Pinto dos Santos & Baeßler, 2018; Pinto dos Santos et al., 2019). Structured reporting may also seamlessly integrate AI results into radiology (European Society of Radiology (ESR), 2023), while natural language processing (NLP) has the potential to accelerate the adoption of structured reporting (Pinto dos Santos & Baeßler, 2018; Pinto dos Santos et al., 2019).

Although most radiologists recognize the numerous advantages of structured reporting, its adoption in clinical practice remains quite limited (Jorg et al., 2023). A survey conducted among members of the Italian Society of Medical Radiology (Faggioni et al., 2017) revealed that 56% of respondents had never used structured reporting in their work. In practical applications, structured reporting faces several challenges, including the risk of oversimplifying complex cases (Faggioni et al., 2017), overly rigid report templates (Faggioni et al., 2017), and a lack of willingness among radiologists to change their established reporting habits (Bergomi et al., 2024; Dos Santos et al., 2019; Faggioni et al., 2017; Haroun et al., 2019; Jorg et al., 2023; Schoeppe et al., 2018). One of the major limitations of structured reporting is its unsuitability for complex cases (Rocha et al., 2020). When dealing with such cases, radiologists may be forced to use templates that are not entirely appropriate, which can restrict the ability to accurately express findings (Olthof et al., 2020; Yousem, 2019). In contrast, free-text reporting offers greater flexibility, allowing radiologists to describe imaging findings in more detail.

Regarding the impact of reporting format on information extraction accuracy and efficiency, studies have found no significant differences between free-text reporting and structured reporting in terms of score, time, and efficiency (Sistrom & Honeyman-Buck, 2005). This suggests that both formats are similarly accurate and effective in conveying case-specific information. Furthermore, a study on head MRI reports for patients with suspected stroke (Johnson et al., 2010) indicated that structured reporting neither improved nor worsened attending physicians' perceptions of report clarity, implying that in certain contexts, the benefits of structured reporting may not be as pronounced.

Additionally, the study (Dos Santos et al., 2019) suggests that structured reporting is highly valuable for preoperative staging. However, in the early postoperative period, where imaging findings are more variable, structured reporting becomes less applicable, whereas free-text reporting proves more effective. Another major barrier to the adoption of structured reporting is radiologists' personal writing preferences. Some radiologists argue that structured reporting limits their ability to express findings in a personalized manner and restricts their stylistic freedom, leading them to favor free-text reporting (Cramer et al., 2014; Haroun et al., 2019).

Beyond these issues related to usability and individual preferences, the implementation of structured reporting also involves significant challenges in terms of template development, testing, and validation (Pesapane et al., 2023), which require substantial time and resources. Moreover, predefined templates that are not regularly updated or adjusted may introduce inaccuracies, further affecting the reliability and clinical applicability of structured reporting (Haroun et al., 2019). Table 5 summarizes the main advantages and disadvantages of structured reporting and free-text reporting.

In ARRG research, most generated reports adopt a free-text reporting format, likely because publicly available datasets predominantly provide free-text reports (Section 4 will introduce the experimental datasets). As mentioned earlier, both free-text reporting and structured reporting have their respective strengths and weaknesses. A potential future solution could be the adoption of a hybrid approach that com-

Table 4
Mammography report in structured format (based on BI-RADS D'Orsi et al., 2013).

UNILATERAL (LEFT) BREAST MAMMOGRAPHY		
Mammography of left breast was performed using low dose radiation technique with adequate compression. Left breast craniocaudal and mediolateral oblique views have been obtained.		
Compared with the previous mammogram dated 15/01/2023, reported as BI-RADS 3, there is an increase in the number of calcifications from the prior exam.		
Clinical presentation:	Non-bloody discharge left breast.	
Breast composition:	ACR Category a	
Finding:		
Mass:	None	
Calcifications:	Location :	Lateral portion on CC view; Superior portion on MLO view
	Suspicious:	Coarse heterogeneous
	Distribution:	Regional
Architectural distortion:	None	
Asymmetries:	None	
Intramammary lymph node:	None	
Skin lesion:	None	
Solitary dilated duct:	None	
Associated features:	None	
Impression:	BI-RADS® Category 3: Probably Benign Finding.	
Advice:	Short interval follow-up of the left breast is recommended in 6 months.	

bines the advantages of both methods. For instance, flexible natural language could be used for complex descriptions, while structured templates could be employed for specific key sections, such as detailed descriptions of lesions.

From an algorithmic perspective, the choice of report format also changes the learning problem. Free-text reports provide richer linguistic variability, but they increase output diversity and make supervision and automatic evaluation more challenging because clinically similar reports may differ substantially in lexical form. In such settings, surface-overlap metrics such as BLEU or ROUGE may become less informative, since reports that are clinically similar can still differ markedly in wording. Structured reports reduce this variability and make supervision, slot-level evaluation, and factual checking easier, although they may oversimplify complex cases and constrain expressive description. This is one reason why a hybrid reporting target may be attractive for ARRG: it offers structured anchors for clinically critical content while preserving enough natural language flexibility to describe atypical findings (Delbrouck et al., 2025; Li et al., 2025; Zhao et al., 2024).

4. Datasets

In the research of ARRG, finding suitable datasets is a critical task. Ideal experimental datasets should include high-quality images that are clear and meet the technical requirements of medical imaging examinations. For instance, a qualified chest X-ray (Chand et al., 2013) should fully display both lung apices and costophrenic angles, with the clavicle medial edges equidistant from the vertebral column to ensure there is no rotation. Moreover, the X-ray should have adequate penetration to faintly visualize the intervertebral discs below T9 and clearly show at least six anterior ribs and ten posterior ribs above the right hemidiaphragm, indicating sufficient inhalation. The medial edges of the scapulae should be positioned outside the lung fields to avoid obscuring critical anatomical structures. These stringent requirements collectively ensure the diagnostic value and quality of the chest X-ray images. In addition, an ideal dataset should come with complete and standardized diagnostic reports. These reports must be detailed and accurate, while adhering to standardized medical documentation formats. However, datasets meeting these criteria are rare in practice. Ethical considerations and patient privacy protection are major obstacles, with many medical institutions exercising caution in data sharing and imposing strict restrictions on access. Additionally, the cost of annotating high-quality diagnostic reports is significant, requiring professional expertise, which further limits the availability of comprehensive datasets.

While datasets containing only medical images (sometimes accompanied by their corresponding segmentation masks) are relatively accessible on open platforms like Kaggle and HuggingFace, and have contributed to foundational medical imaging tasks such as segmentation and classification (Ehab et al., 2024; Huang et al., 2024), datasets required for ARRG are more complex. They typically necessitate the integration of raw medical images, and detailed diagnostic reports. Such multimodal datasets are extremely scarce in the public domain, with most remaining either non-public or under restricted access. This

Table 5
A summary of the advantages and disadvantages of free-text reporting and structured reporting.

Reporting format	Advantages	Disadvantages
Free-text Reporting	1. High flexibility: Suitable for detailed descriptions of complex cases, allowing adjustments based on specific situations. 2. Rich expression: Allows the use of natural language for more detailed and personalized descriptions. 3. Quick recording: Doctors can quickly document observations without following a fixed format.	1. Inconsistent language: Varied descriptive styles and language among doctors may lead to ambiguity or misinterpretation. 2. Lack of standardization: Difficult to systematically analyse, hindering data mining and statistical analysis. 3. Potential omission of critical information: Important details, such as preoperative tumour staging, may be missed. 4. Not conducive to automated processing: Difficult to integrate with AI or other automated tools.
Structured Reporting	1. High completeness and comparability: Ensures key information is not omitted, facilitating comparisons. 2. Reduced ambiguity: Standardized language and format improve communication efficiency with clinicians. 3. Improved consistency and reproducibility: Reports are more standardized, facilitating follow-up research and analysis. 4. Facilitates data extraction and mining: Suitable for research and education. 5. Promotes AI development: Structured data is more suitable as training and validation datasets for AI.	1. Lack of flexibility: Fixed templates may not suit complex or special cases. 2. Limits personalized expression: Doctors' descriptions may be constrained by templates, unable to fully express details. 3. Learning curve: Doctors may need extra time to learn and adapt. 4. Template updates may lag: Failure to update templates promptly may result in outdated or inaccurate content.

scarcity not only limits researchers' freedom to develop models but also hinders the further advancement and practical application of ARRG technology.

This section will enumerate the datasets commonly used in ARRG research, including publicly available, restricted-access, and private datasets. These frequently utilized experimental datasets encompass both two-dimensional (2D) and three-dimensional (3D) data. It is worth noting that this review focuses specifically on radiology-related medical data; thus, the datasets discussed primarily include those involving X-rays, CT, MRI, PET-CT and ultrasound (US) imaging. Medical data from other fields, such as pathology or optometry, are not included in the scope of this literature review. Additionally, this section will highlight certain datasets frequently mentioned in ARRG literature reviews, which, when used independently, are not fully suitable for ARRG tasks. For instance, datasets like ChestX-ray8 dataset (Wang et al., 2017), ChestX-ray14 (Wang et al., 2017), and CheXpert dataset (Irvin et al., 2019) only provide disease labels without accompanying diagnostic reports. These datasets are typically employed for model pretraining (Li et al., 2019a,b, 2018; Rodin et al., 2019; Xiong et al., 2019; Yuan et al., 2019; Zhang et al., 2020) or fine-tuning (Jing et al., 2020), or are used in conjunction with other datasets that include diagnostic reports (Wang et al., 2018; Xue & Huang, 2019). The primary reason is that an ideal dataset for ARRG tasks not only requires high-quality medical images but also necessitates comprehensive and standardized diagnostic reports. Without such reports, these datasets alone cannot fully satisfy the requirements of ARRG tasks. Nonetheless, these datasets remain valuable for research purposes. Their large scale and rich annotations make them particularly useful for learning generalizable medical imaging features during pretraining. Subsequently, models can be fine-tuned with datasets that include complete diagnostic reports to enhance the quality of generated outputs for ARRG tasks.

4.1. Public datasets

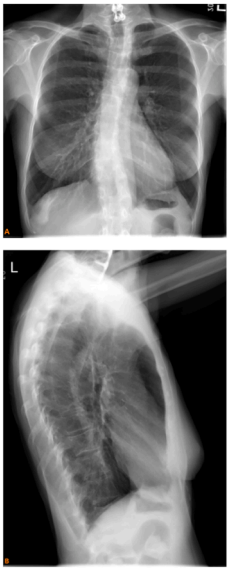
Public datasets refer to datasets that are freely accessible and downloadable by any researcher or the general public without requiring special permissions or institutional approval.

*4.1.1. **Indiana University X-ray (IU X-ray dataset)** (Demner-Fushman et al., 2016)*

The Indiana University X-ray Dataset (Demner-Fushman et al., 2016) was released in 2016. It is also known as the OpenI dataset. This publicly available dataset was collected by Indiana University, USA, and consists of 7470 frontal or lateral chest X-rays and 3955 corresponding radiology reports. Each radiology report is divided into four sections: Comparison, Indication, Findings, and Impression, which are largely aligned with the ESR guidelines. However, the reports are unstructured presented as free text. The official source does not provide a clear data splitting method. Some studies (Chen et al., 2022; Huang et al., 2023; Liu et al., 2021a; Wang et al., 2023b) have adopted the data splitting method proposed by Chen et al. to ensure reproducibility (Chen et al., 2020). Furthermore, the dataset includes manually annotated Medical Subject Headings (MeSH) (FB, 1963) and RadLex (Langlotz, 2006) terms, along with automated annotations of MeSH terms (FB, 1963) and negation information using the Medical Text Indexer (MTI) (Mork et al., 2013) and MetaMap (Aronson & Lang, 2010) tools. Fig. 3 is an example of an X-ray and report pair from the IU X-ray dataset (Demner-Fushman et al., 2016). The top-left image is a frontal chest X-ray, the bottom-left image is a lateral chest X-ray, and the right part is the corresponding diagnostic report and MeSH (FB, 1963) terms.

In the context of literature reviews, the IU X-ray dataset (Demner-Fushman et al., 2016) is one of the most commonly used datasets. Its widespread usage is primarily due to its public accessibility, as well as the pairing of the all chest X-rays with corresponding radiology reports, making it ideal for research on image-to-text tasks. Additionally, the dataset's moderate size reduces computational requirements, making it suitable for early-stage model design and training.

Despite its advantages, the IU X-ray dataset (Demner-Fushman et al., 2016) has several limitations due to its scale. First, the dataset is relatively small. Compared to larger datasets, such as MIMIC-CXR (Johnson et al., 2019a,b) and PadChest dataset (Bustos et al., 2020), the smaller size of the IU X-ray dataset (Demner-Fushman et al., 2016) may restrict the generalizability of models. Some people (Chen et al., 2022, 2020) suggested that dataset size significantly impacts model performance. To address the issue of insufficient data, pre-trained models or data augmentation techniques can be utilized (Chen et al., 2020). Second, the dataset only includes a single examination record for each patient, making it unsuitable for supporting longitudinal studies, such as follow-ups or comparative analyses across multiple time points. Finally, the dataset is primarily derived from outpatient examinations, lacking chest X-rays from hospitalized patients, such as those showing central venous catheters or similar projections. This limitation may affect the accuracy and generalizability of models trained on this dataset (Sloan et al., 2024).



Indication: Shortness of breath

Comparison: None.

Findings: The cardiac contours are normal. The lungs are hyperinflated with flattening of the diaphragms and tapering of the distal pulmonary vasculature. There is no focal consolidation. Thoracic spondylosis. Mild dextroscoliosis of the spine. Prior anterior cervical fusion.

Impression: Emphysema without superimposed pneumonia.

MeSH:

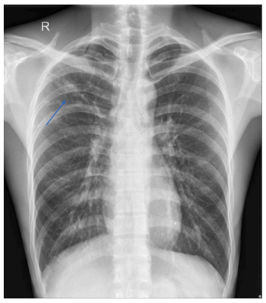
Lung/hyperdistention;Diaphragm/flattened;Spondylosis/thoracic vertebrae;Scoliosis/right/mild;Spinal Fusion/cervical vertebrae/anterior;Emphysema

Fig. 3. This is an example of an X-ray and report pair from the IU X-ray dataset (Demner-Fushman et al., 2016). The top-left image is a frontal chest X-ray, the bottom-left image is a lateral chest X-ray, and the right part is the corresponding diagnostic report and MeSH (FB, 1963) terms.

4.1.2. *China set - The Shenzhen set - Chest X-ray database (Candemir et al., 2013; Jaeger et al., 2013)*

The Shenzhen Chest X-ray Dataset (Candemir et al., 2013; Jaeger et al., 2013) was jointly created by the Shenzhen No. 3 People's Hospital, Guangdong Medical College, Shenzhen, China, and the National Library of Medicine, Maryland, USA. The dataset contains chest X-ray images sourced from outpatient clinics, including 336 cases showing tuberculosis manifestations and 326 normal cases, totaling 662 images. All images are in PNG format with a resolution of approximately 3K × 3K, providing clear quality that fully meets the standards of high-quality chest X-rays. The dataset consists of two parts: chest X-ray images and corresponding clinical readings documents. The documents record basic patient information (such as gender and age) as well as diagnostic results, including the location of lesions and the type of tuberculosis. The clinical readings documents contain information about lesion locations, making it possible to transform them into radiology reports or even structured reports. Combined with the dataset's clear, high-quality images that meet the requirements of qualified chest X-rays (Chand et al., 2013) and the use of easily viewable and processable PNG format, the Shenzhen Chest X-ray dataset (Candemir et al., 2013; Jaeger et al., 2013) has the potential to be an ideal research dataset for ARRG. However, this dataset also has some limitations. First, the dataset is relatively small. ARRG research typically requires large-scale paired image-text datasets for model training. A previous study (Chen et al., 2020) has shown that the size of a dataset affects model performance. Compared to commonly used ARRG datasets, such as MIMIC-CXR (Johnson et al., 2019a,b), the Shenzhen dataset is Candemir et al. (2013) and Jaeger et al. (2013) relatively small, which may limit its effectiveness when used independently. Second, the dataset focuses on a single disease type. The Shenzhen dataset (Candemir et al., 2013; Jaeger et al., 2013) primarily targets tuberculosis and does not include data for other pulmonary diseases (e.g., pneumonia, tumours). In contrast, ARRG research often requires diverse disease types to improve the generalizability of models.

Given these limitations, it may be beneficial to combine the Shenzhen dataset (Candemir et al., 2013; Jaeger et al., 2013) with other datasets (e.g., IU X-ray dataset Demner-Fushman et al., 2016, MIMIC-CXR dataset Johnson et al., 2019a,b). Specific approaches could include using the Shenzhen Chest X-ray dataset (Candemir et al., 2013; Jaeger et al., 2013) for model fine-tuning or pretraining, while leveraging other



Male , 32yrs secondary PTB in the right upper field

Fig. 4. This is a chest X-ray image from the Shenzhen Chest X-ray dataset (Candemir et al., 2013; Jaeger et al., 2013). The corresponding clinical reading states "Male, 32 years old, secondary PTB in the right upper field". The lesion is indicated with blue arrows.

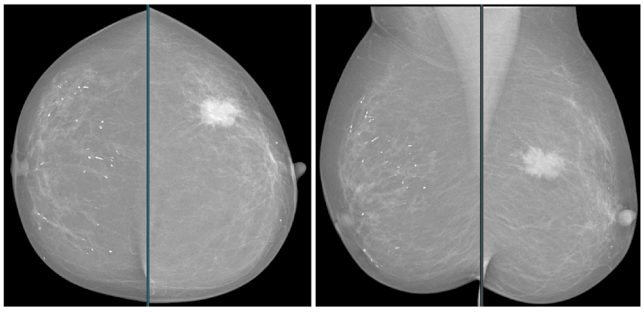
datasets for validation. For example, the high-quality images and precise lesion information in the Shenzhen dataset (Candemir et al., 2013; Jaeger et al., 2013) could be used to train models, which can then be evaluated and validated using larger and more diverse datasets, such as IU X-ray dataset (Demner-Fushman et al., 2016) or MIMIC-CXR dataset (Johnson et al., 2019a,b). Fig. 4 is a chest X-ray image from the Shenzhen Chest X-ray dataset (Candemir et al., 2013; Jaeger et al., 2013). The corresponding clinical reading states: "Male, 32 years; secondary PTB in the right upper field". The lesion is indicated with blue arrows.

4.1.3. *INbreast dataset (Moreira et al., 2012)*

The INbreast dataset (Moreira et al., 2012) is provided by a breast center located in a university hospital (Centro Hospitalar de S. Jo ao [CHSJ], Breast Centre, Porto). It contains 115 cases with a total of 410 images, including 90 cases from patients with bilateral breast involvement (four images per case) and 25 cases from mastectomy patients (two images per case). The dataset includes screening, diagnostic, and follow-up mammography images. The image types cover normal cases, masses, calcifications, architectural distortions, asymmetries, and various combinations of lesions. The annotations in INbreast dataset (Moreira et al., 2012) are performed and validated by radiology experts, providing detailed lesion contour information in XML format. Additionally, the dataset includes patient information such as age, breast density (based on ACR standards), and BI-RADS (D'Orsi et al., 2013) classification. Biopsy results are also available for some cases. This dataset has been applied in ARRG research (Sun et al., 2019). Fig. 5 is an example from the INbreast dataset (Moreira et al., 2012). The top-left image shows the CC view of both breasts, while the top-right image shows the MLO view of both breasts. The text below is the corresponding diagnostic report, originally written in Portuguese and translated into English as: The imaging study documented a nodule located in the upper outer quadrant (QSE) of the left breast, measuring 6 cm in diameter. An ultrasound-guided core biopsy was performed, collecting 4 fragments for anatomical-pathological analysis. A fine-needle aspiration biopsy of the axillary lymph node was also performed. Preoperative marking with carbon was carried out. Imaging findings are highly suggestive of malignancy - Bi-RADS 5.

4.1.4. *Lumbar spine MRI dataset (Al-Kafri et al., 2019; Sudirman et al., 2019a)*

The Lumbar Spine MRI dataset (Al-Kafri et al., 2019; Sudirman et al., 2019a) originates from Irbid Specialty Hospital in Jordan and includes MRI scans of 515 patients (initially 575 cases were collected, but 515 were retained after data cleaning). After downloading the data (Sudirman et al., 2019a), there are 575 3D MRI datasets and 575 corresponding diagnostic reports available. The dataset features T1-weighted and T2-weighted images with sagittal and axial views. It includes annotations by radiologists identifying pathological regions and diagnostic information (Sudirman et al., 2019b), such as intervertebral disc protru-



Nódulo sólido com cerca de 4 cm de maior diâmetro, suspeito, mama esquerda, QSE. Aderência cutânea.

Fez MB - 5 fragmentos.

Fez BA a gânglio de aspecto patológico.

Fig. 5. This is an example from the INbreast dataset (Moreira et al., 2012). The top-left image shows the CC view of both breasts, while the top-right image shows the MLO view of both breasts. The text below is the corresponding diagnostic report, originally written in Portuguese.

sion, thecal sac compression, central or foraminal stenosis, and endplate degeneration, among other conditions. This dataset comprises multi-slice scans, exhibiting 3D characteristics, and serves as a valuable resource for studying lumbar spine-related pathologies. Fig. 6 represents a case from the Lumbar Spine MRI dataset (Al-Kafri et al., 2019; Sudirman et al., 2019a). From left to right and top to bottom, it includes a localizer image, two T2 axial images, two T1 axial images, a T2 sagittal image, a T1 sagittal image, and the corresponding radiological report.

4.1.5. *COVID-19 CT report (COV-CTR) dataset (Li et al., 2023b)*

The COVID-19 CT Report (COV-CTR) dataset (Li et al., 2023b) is a public medical database derived from the public COVID-CT dataset (Yang et al., 2020). It consists of lung CT scan images (2D data) and their corresponding diagnostic reports, which are available in both Chinese and English. The dataset includes 728 images, with 349 classified as COVID-19 positive cases and 379 as non-COVID-19 cases. These images were sourced from published research papers and were analysed by three Chinese radiologists with over five years of experience, who generated the diagnostic reports. Additionally, the dataset includes 68 medical tags, comprising 50 abnormal tags and 18 normal tags. Examples of studies that have applied this dataset include (Li et al., 2023b; Song et al., 2024; Wang et al., 2021; Zhang et al., 2024a, 2023a). Fig. 7 is an example from the COV-CTR dataset (Li et al., 2023b). From top to bottom, it displays a lung CT image, along with the corresponding Chinese and English reports.

4.1.6. *CTRG-Brain-263K dataset (Tang et al., 2024)*

CTRG-Brain-263K is a dataset dedicated to brain CT images, which was released by Tang et al., comprising 263,670 brain CT scan images and 10,009 diagnostic reports written by radiology experts (Tang et al., 2024). Among these cases, 6007 are abnormal, while 4002 are normal, with an abnormal-to-normal ratio of approximately 1.5:1. The reports in this dataset were written and reviewed by eight professional radiologists. This dataset includes fully detailed Chinese reports, which were initially translated using the Baidu Medical Translation API and subsequently manually reviewed to ensure high-quality bilingual (Chinese-English) versions. Additionally, the dataset covers seven key medical observation categories: cerebral hemispheres, brain parenchyma, midline structures, sulcus and fissures, brainstem and cerebellum, bone window, and maxillary sinus. Fig. 8 provides an example from the CTRG-Brain-263K dataset (Tang et al., 2024). From top to bottom, it displays brain CT images, followed by the corresponding Chinese and English reports.

4.1.7. *CTRG-Chest-548K dataset (Tang et al., 2024)*

CTRG-Chest-548K is a large-scale chest CT image report generation dataset proposed by Tang et al., comprising 548,696 chest CT images and 1804 diagnostic reports, with each case typically containing multiple CT scan images (Tang et al., 2024). This dataset covers a wide range of pathological conditions, from mild abnormalities to severe lesions, while also including a small number of completely normal cases. Similar to the CTRG-Brain-263K dataset (Tang et al., 2024), all diagnostic reports in this dataset were written and reviewed by eight professional radiologists. The reports were then translated using the Baidu Medical Translation API and manually verified to ensure high-quality and accurate medical terminology and descriptions. The CTRG-Chest-548K dataset (Tang et al., 2024) encompasses eight major medical observation categories, including thoracic structures, ribs, lung window, heart, pleura, liver, kidneys, and thyroid.

4.1.8. *Pathology education informational resource (PEIR) digital library dataset (PEIR Digital Library, 2025)*

The Pathology Education Informational Resource (PEIR) Digital Library dataset (PEIR Digital Library, 2025) contains a vast array of medical data, including pathological and radiological information. The radiological section currently features a collection of 4732 selected radiology teaching images, copyrighted by the Department of Pathology at the University of Alabama at Birmingham. This collection encompasses imaging of 20 major human organs and systems, with modalities including ultrasound, X-ray, CT, and MRI. Each case typically consists of several images, accompanied by sentence-level descriptions that include diagnostic information and, in some instances, patient history and clinical details. In summary, each medical imaging image is accompanied by a sentence-level description. Fig. 9 illustrates seven contrast-enhanced CT images from the same case within the PEIR Digital Library dataset (PEIR Digital Library, 2025), along with their corresponding descriptions.

4.1.9. *ImageCLEF 2015 liver CT annotation dataset (Marvasti et al., 2015)*

The ImageCLEF 2015 Liver CT Annotation dataset (Marvasti et al., 2015) was designed for a semantic annotation challenge focused on liver CT images (3D). It includes 50 training datasets and 10 test datasets, and corresponding structured reports. Each dataset contains a cropped CT volume of the liver, a liver mask defining the liver region, a region of interest (ROI) marking the lesion area, and 73 semantic features (UsE) annotated based on the ONLIRA ontology. The goal of the task is to generate a standardized radiology report composed of UsE features using the cropped liver CT volume and the LiCO (Liver Case Ontology). The relevant literature indicates that this dataset is a public dataset; however, during the process of writing this literature review, no accessible source for obtaining the dataset was found. Nonetheless, since the dataset was used in some studies (Loveymi et al., 2020, 2021), it is still considered a public dataset.

4.1.10. *Radiology objects in COntext (ROCO) dataset (Pelka et al., 2018)*

The ROCO dataset, created by Pelka et al., was derived from the PubMed Central Open Access subset (Pelka et al., 2018). It includes 81,825 radiological images and over 6127 non-radiological images, covering a wide range of medical imaging modalities such as CT, MRI, X-ray, ultrasound, and PET. Each image in the dataset is accompanied by detailed captions, and preprocessed to extract keywords and medical semantic information, including Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) and Semantic Types (SemTypes). Fig. 10 presents an example from the ROCO dataset (Pelka et al., 2018), featuring a cardiac color Doppler ultrasound image along with its associated medical semantic information.

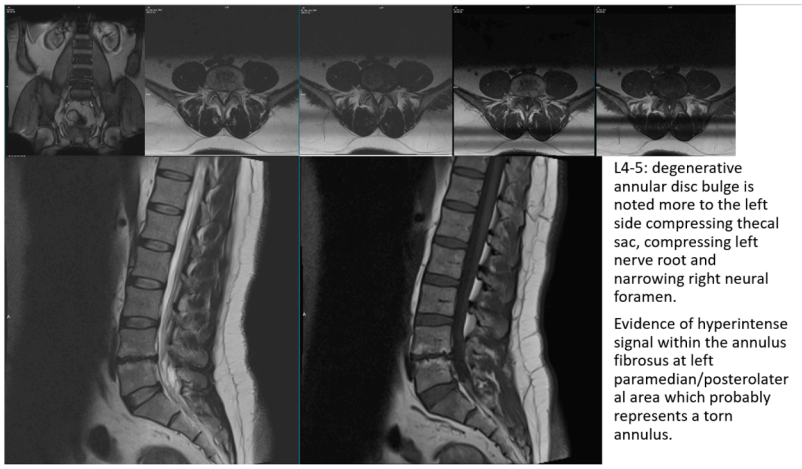
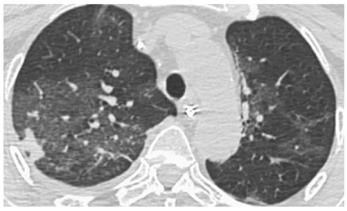


Fig. 6. It represents a case from the Lumbar Spine MRI dataset (Al-Kafri et al., 2019; Sudirman et al., 2019a). From left to right and top to bottom, it includes a localizer image, two T2 axial images, two T1 axial images, a T2 sagittal image, a T1 sagittal image, and the corresponding radiological report.



Findings:

胸廓对称，纵隔心影居中，纵隔内未见肿大淋巴结。双肺上叶可见片状磨玻璃影，边缘不清，右肺上叶胸膜下见斑片状实变影。所见叶段支气管通畅，双侧胸腔内未见异常密度影。

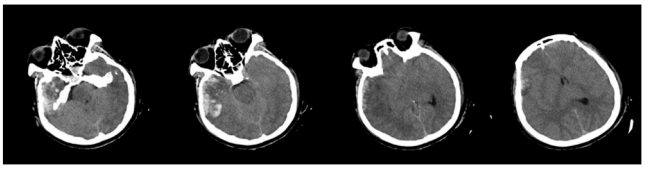
The thorax was symmetrical, the mediastinal heart shadow was centered, and no enlarged lymph nodes were seen in the mediastinum. The upper lobe of both lungs was seen as a lamellar ground glass shadow with indistinct margins, and a patchy solid shadow was seen under the pleura of the upper lobe of the right lung. The bronchi of the lobes were clear, and no abnormal density shadow was seen in the bilateral thorax.

Impression:

双肺上叶炎性病变（符合病毒性肺炎）

Inflammatory lesions in the upper lobes of both lungs (consistent with viral pneumonia). (This sentence in the Impression section was translated by the author from the original Chinese text.)

Fig. 7. An example from the COV-CTR dataset (Li et al., 2023b). From top to bottom, it displays a lung CT image, along with the corresponding Chinese and English reports.



Report in Chinese:

CT所见： 右侧额颞叶见斑片状高密度影，周围见低密度水肿带，右侧额颞部颅板下见条带状高密度影，邻近脑沟及外侧裂池密度增高；余各脑室、脑池大小形态正常，中线结构居中，幕下小脑、脑干无异常。骨窗示左侧额骨及颅底见线样低密度影，左侧顶部皮下软组织肿胀、积气，两侧筛窦及蝶窦内积液。
CT印象： 右侧额叶、颞叶脑挫伤，右侧额颞部硬膜下血肿，蛛网膜下腔出血。左侧颞骨及颅底骨折。左侧顶部皮下软组织肿胀、积气，两侧筛窦及蝶窦内积液。

Report in English:

CT Findings: Patchy high density shadow is seen in the right frontal lobe, low-density edema zone is seen around, banded high density shadow is seen under the right frontal temporal cranial plate, and the density of adjacent sulcus and lateral fissure cistern is increased: the size and morphology of the remaining ventricles and cisterns are normal, the midline structure is in the middle, and the infratentorial cerebellum and brain stem are normal. The bone window shows a linear low density shadow in the left holoskeleton and skull base, swelling and gas accumulation in the subcutaneous soft tissue at the top of the left side, and fluid accumulation in both ethmoid and sphenoid sinuses.
CT Impression: Brain contusion in the right frontal lobe and temporal lobe, subdural hematoma in the right frontotemporal region, subarachnoid hemorrhage. Left temporal bone and skull base fracture. The subcutaneous soft tissue at the top of the left side swelled and accumulated air, and fluid accumulated in both ethmoid sinuses and sphenoid sinuses.

Fig. 8. An example from the CTRG-Brain-263K dataset (Tang et al., 2024). From top to bottom, it displays brain CT images, followed by the corresponding Chinese and English reports..

4.1.11. *Radiology object in COntext version 2 (ROCOv2) dataset*

The ROCOv2 dataset (Rückert et al., 2024) is a multimodal dataset comprising 79,789 radiological images extracted from the PubMed Open Access subset, along with detailed captions and associated medical semantic information. Compared to its predecessor, the ROCO dataset, ROCOv2 represents an updated version that includes 35,705 new images added to PubMed since 2018. Additionally, it features manually curated medical concepts, including clinical modalities, anatomy (X-rays), and directionality (X-rays). Fig. 11 shows an example from the ROCOv2 dataset (Rückert et al., 2024), featuring a whole-body F18-FDG PET/CT scan showing multiple enlarged lymph nodes in the left supraclavicular area.

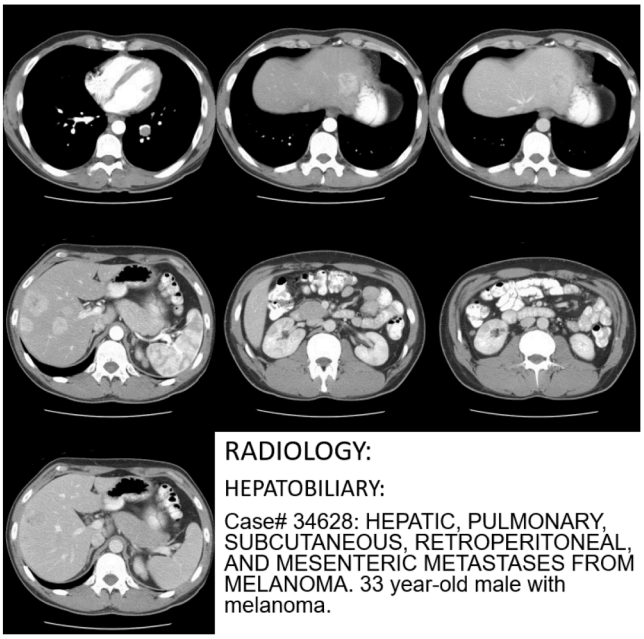
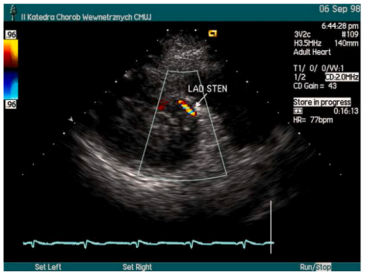


Fig. 9. It is an example from PEIR (PEIR Digital Library, 2025), which consists of seven contrast-enhanced CT images of the same case and their corresponding descriptions (free-text reporting).

4.1.12. *Rad-ReStruct dataset (Pellegrini et al., 2023)*

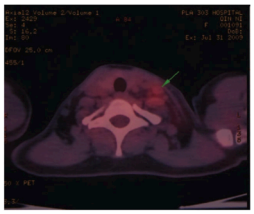
The Rad-ReStruct dataset (Pellegrini et al., 2023), developed by Pellegrin et al. based on the IU X-Ray dataset (Demner-Fushman et al., 2016), aims to advance research in structured radiology reporting. This dataset comprises 3720 images and 3597 structured reports annotated in a fine-grained and hierarchically organized manner, containing a large number of questions. Pellegrin et al. modeled the structured reporting task as a hierarchical Visual Question Answering (VQA) task. Its hierar-



Colour Doppler

echocardiography, parasternal modified short axis view. Middle LAD stenosis was incidentally found on routine TTE. It was later confirmed by coronary angiography. See movie 13 [see Additional file 13]

Fig. 10. It presents an example from the ROCO dataset (Pelka et al., 2018), featuring a cardiac color Doppler ultrasound image along with its associated medical semantic information.



PET - CT

The whole body F18-FDG PET/CT tomography scan (July 31, 2009) showing multiple enlarged lymph nodes on left supraclavicular area.

Fig. 11. It presents an example from the ROCOv2 dataset (Rückert et al., 2024), featuring a whole-body F18-FDG PET/CT scan showing multiple enlarged lymph nodes in the left supraclavicular area.

chical reporting template consists of three levels, focusing respectively on the presence of abnormalities (e.g., diseases, findings, or abnormal regions), specific elements (e.g., particular diseases), and detailed attributes (e.g., severity or location).

4.1.13. *CheXpert dataset (Irvin et al., 2019)*

The CheXpert dataset (Irvin et al., 2019) is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs collected from 65,240 patients at Stanford Hospital. This dataset provides 14 chest disease labels, including No Finding, Enlarged Cardiom., Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices. Each label is annotated as positive, negative, or uncertain based on the interpretation of the radiographs.

Although the CheXpert dataset (Irvin et al., 2019) does not contain radiology diagnostic reports, it can be combined with other datasets, such as the IU X-ray dataset (Demner-Fushman et al., 2016), which includes complete radiology reports. The CheXpert dataset (Irvin et al., 2019) can play a significant role in ARRG research in several ways. First, as a large-scale dataset with labeled data, it can be used as pretraining data (Zhang et al., 2020). Second, it can be used for disease classification tasks (Yuan et al., 2019). The CheXpert dataset (Irvin et al., 2019) dataset provides 14 chest disease labels, making it a suitable choice for training multi-label disease classification models. Third, it can provide prior knowledge for constructing chest abnormality graphs (Zhang et al., 2020). Lastly, other datasets that include radiology diagnostic reports, such as the IU X-ray dataset (Demner-Fushman et al., 2016), can be used for fine-tuning.

4.1.14. *CheXpert plus dataset (Chambon et al., 2024; Medicine, 2024)*

CheXpert Plus (Chambon et al., 2024; Medicine, 2024) is an extended version of the original CheXpert dataset (Irvin et al., 2019), expanding it from a conventional chest radiograph classification dataset into a large-scale multimodal resource that integrates chest X-ray images, radiology reports, DICOM metadata, patient demographic information, pathology labels, and RadGraph annotations. According to the official release, the dataset contains 223,462 paired chest X-rays and radiology reports from 187,711 studies involving 64,725 patients. The reports are divided into up to 11 subsections, such as Findings, Impression, History, Comparison, and Technique, making the dataset more suitable for report-level and section-aware modeling. In addition, CheXpert Plus (Chambon et al., 2024; Medicine, 2024) provides labels for 14 chest pathologies, supporting not only ARRG research but also multimodal pretraining, medical information extraction, and clinically oriented evaluation.

4.1.15. *ChestX-ray8 dataset & ChestX-ray14 dataset (Wang et al., 2017)*

The ChestX-ray8 dataset (Wang et al., 2017) is a publicly available dataset released by the National Institutes of Health (NIH). It contains 108,948 frontal chest X-ray images collected from 32,717 unique patients. This dataset includes eight common thoracic pathology keywords, covering the following common thoracic diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. Additionally, images without any detected pathological findings are labeled as “Normal”.

The ChestX-ray14 dataset (Wang et al., 2017) is a publicly available dataset released by the National Institutes of Health (NIH). It is an extended version of the ChestX-ray8 dataset (Wang et al., 2017). Building upon the original 8 thoracic diseases, it adds 6 new pathologies (Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia), covering a total of 14 common thoracic diseases. Compared to ChestX-ray8 dataset (Wang et al., 2017), the ChestX-ray14 dataset (Wang et al., 2017) has been significantly expanded and includes 112,120 frontal chest X-ray images.

Similar to the CheXpert dataset (Irvin et al., 2019), ChestX-ray 8 (Wang et al., 2017) and ChestX-ray14 (Wang et al., 2017) are also frequently used in conjunction with other datasets containing complete radiology reports, such as IU X-ray (Demner-Fushman et al., 2016) or MIMIC-CXR datasets (Johnson et al., 2019a,b). These datasets are typically employed for model pretraining, fine-tuning, or disease classification experiments. Subsequently, datasets with complete radiology reports (e.g., IU X-ray Demner-Fushman et al., 2016 or MIMIC-CXR Johnson et al., 2019a,b) are integrated to advance research on ARRG. For example, some studies (Biswal et al., 2020; Li et al., 2019a,b, 2018) use ChestX-ray8 (Wang et al., 2017) for pretraining; similarly, other studies utilize ChestX-ray14 (Wang et al., 2017) for pretraining (Xiong et al., 2019), fine-tuning (Jing et al., 2020), and disease classification (Wang et al., 2018; Xue & Huang, 2019), followed by the incorporation of datasets with complete radiology reports to carry out ARRG research.

4.1.16. *DeepLesion dataset (Yan et al., 2018)*

DeepLesion dataset (Yan et al., 2018) is a large-scale medical imaging dataset released by the Clinical Center of the National Institutes of Health (NIHCC). The dataset contains 32,120 axial CT slices from 10,594 CT scans (studies) of 4427 independent patients. Each image contains 1 to 3 lesions, with bounding boxes and size measurements, totaling 32,735 lesions. The dataset covers various regions of the body, including the lungs, liver, bones, and more. Each lesion is meticulously annotated with information such as type, location, size, and bounding box provided by radiologists. The lesions in DeepLesion (Yan et al., 2018) are highly diverse, including both benign and malignant tumours, cysts, inflammations, and lesions of various sizes, shapes, and appearances. It also includes challenging cases such as small lesions, low-contrast lesions, and lesions located in complex anatomical structures. While the dataset does not contain full radiology reports, it can be used for pretraining, fine-tuning, and other tasks in studies like those of ARRG.

4.2. *Restricted dataset*

Restricted datasets are datasets that require researchers to submit an application and obtain authorization before gaining access. These datasets are typically managed by medical institutions or research organizations and require applicants to comply with ethical approvals or data use agreements.

4.2.1. *Medical information mart for intensive care chest X-Ray (MIMIC-CXR dataset) (Johnson et al., 2019a,b)*

The Medical Information Mart for Intensive Care Chest X-Ray (MIMIC-CXR) (Johnson et al., 2019a,b) is currently one of the largest ARRG datasets. It includes 377,110 chest X-ray images and 227,835 corresponding unstructured textual reports. The data was collected between 2011 and 2016, including 64,588 patients from the Beth Israel Deaconess Medical Center, primarily from its emergency department services in Boston, MA. The initial release of MIMIC-CXR dataset (Johnson et al., 2019a,b) was in DICOM format, but for ease of use, the images were later converted to JPEG format, resulting in the MIMIC-CXR-JPG dataset (Johnson et al., 2019b), which contains 377,110 chest X-ray images and 227,827 corresponding unstructured textual reports. In some literature reviews, MIMIC-CXR dataset (Johnson et al., 2019a,b) and MIMIC-CXR-JPG dataset (Johnson et al., 2019a,b) are treated as the same dataset and collectively referred to as MIMIC-CXR, while in others, they are analysed separately. In this review, we consider these two datasets as a single dataset and refer to them collectively as MIMIC-CXR (Johnson et al., 2019a,b). This dataset also includes MeSH (FB, 1963) to facilitate tag-based research. It provides official splits for training, testing, and validation, making it easier to compare the performance of different ARRG models. Researchers wishing to access MIMIC-CXR dataset (Johnson et al., 2019a,b) must register on the PhysioNet platform, complete a CITI training course, and sign a Data Use Agreement (DUA) to ensure ethical use of the data. During the review process, it was found that the MIMIC-CXR dataset (Johnson et al., 2019a,b) is one of the primary datasets used in ARRG research. Its standout advantage lies in being the largest ARRG dataset to date. Some studies (Ying, 2019) suggested that larger datasets often improve model performance. However, accessing the MIMIC-CXR dataset (Johnson et al., 2019a,b) is a relatively complex process, requiring the submission of relevant applications and awaiting approval.

4.2.2. *MIMIC Derivatives dataset*

The MIMIC-CXR dataset (Johnson et al., 2019a,b), known for its extensive data resources and widespread application in research, has inspired the creation of multiple derivative datasets. These datasets are also hosted on the PhysioNet platform. Similar to accessing the MIMIC-CXR dataset (Johnson et al., 2019a,b), researchers must complete a certification process before using these derivative datasets. The research necessity of these derivative datasets lies in the fact that the original MIMIC-CXR corpus, although highly valuable, does not explicitly provide all of the structured signals needed for downstream ARRG development. Derivative resources fill different gaps. Some remove problematic prior-study references that may otherwise encourage hallucinated references to prior studies (Ramesh et al., 2022a,b), some provide region-level grounding and anatomical localization (Krishna et al., 2017; Serra et al., 2023a,b), and some add temporally sensitive annotations (Bannur et al., 2023a,b) or graph-structured supervision (Krishna et al., 2017) that are useful for factual evaluation, disease progression modelling, and multimodal alignment. Accordingly, these derivatives are not merely larger variants of MIMIC-CXR, but resources that add structured signals useful for downstream ARRG-related analysis and modelling. Although only a limited number of ARRG studies have drawn on such derivatives directly, these resources are relevant to tasks involving grounding, temporal information, graph-structured evaluation, and clinically informed report generation. In other words, these derivatives do not merely enlarge MIMIC-CXR; they convert it into a richer ecosystem for pretraining, grounding, error analysis, and clinically informed report generation.

- **MIMIC-ABN Dataset** (Ni et al., 2020): Ni et al. developed a derivative dataset called MIMIC-ABN dataset (Ni et al., 2020), which was created from the MIMIC-CXR dataset (Johnson et al., 2019a,b). MIMIC-ABN dataset (Ni et al., 2020) includes only abnormal images and their corresponding reports from MIMIC-CXR dataset (Johnson et al., 2019a,b), excluding normal images and reports. In other words, every image in the dataset contains at least one abnormality (such as pneumonia, tuberculosis, or pneumothorax) and does not include reports labeled as “no findings. Some studies have utilized this dataset (Hou et al., 2024, 2023a; Liu et al., 2024; Mei et al., 2024; Yan et al., 2021).
- **MIMIC-CXR with Prior References Omitted (CXR-PRO) Dataset (Ramesh et al., 2022a,b):** Ramesh et al. introduced the CXR-PRO dataset (Ramesh et al., 2022a,b), which differs from MIMIC-CXR dataset (Johnson et al., 2019a,b) by employing GILBERT to transform the task of removing references to prior reports into a Named Entity Recognition (NER) task. As a result, the CXR-PRO dataset (Ramesh et al., 2022a,b) eliminates all references to prior reports in the radiology texts, retaining only descriptions relevant to the current image for each report. This prevents the generation model from outputting erroneous references to nonexistent prior reports. The dataset contains 374,139 free-text radiology reports and their corresponding chest X-ray images. Similarly, some studies have also applied this dataset. (Bernardi & Cimitile, 2024; Ranjit et al., 2023).
- **Chest ImaGenome Dataset** (Krishna et al., 2017): The Chest ImaGenome dataset (Krishna et al., 2017), developed by Wu et al. based on the MIMIC-CXR dataset (Wu et al., 2021), was inspired by the Visual Genome effort in the computer vision community (Krishna et al., 2017). This dataset includes annotations for 242,072 chest X-ray images, with each image represented by a scene graph that describes 29 anatomical regions, 1256 combinations of relational annotations, and provides bounding box coordinates and attributes for the anatomical regions. Additionally, the dataset includes over 670,000 localized comparison relationships (e.g., improved, worsened, or no change) between anatomical regions, making it particularly useful for studying temporal trends in sequential imaging exams. As a gold standard, the dataset also provides a manually annotated reference scene graph dataset containing 500 unique patients, supporting model evaluation and benchmarking. The Chest ImaGenome dataset (Krishna et al., 2017) does not include complete radiology reports. However, in the experiments conducted by Serra et al., this dataset was utilized to train a multi-task Faster R-CNN (Ren et al., 2016) model with the primary goal of achieving anatomical localization and finding detection (Serra et al., 2023b). Through this model, the researchers extracted a type of visual feature called “Finding-Aware Anatomical Tokens. These tokens effectively integrate anatomical region information with associated findings and serve as visual features input into a multimodal Transformer model. Combined with textual input (such as clinical indication information), these features are ultimately used to generate radiology reports. In Serra et al.’s another experiment, the Chest ImaGenome dataset (Krishna et al., 2017) was used to provide annotations of anatomical regions in chest X-ray images as well as mappings between report sentences and corresponding anatomical regions (Serra et al., 2023a). It was also utilized to train the Faster R-CNN model (Ren et al., 2016) for extracting visual features of anatomical structures (referred to as anatomical tokens). These annotations supported the implementation of the sentence-anatomy dropout strategy, enabling the model to generate partial reports corresponding to specific anatomical regions based on the input.
- **MS-CXR Dataset** (Boecking et al., 2022): The MS-CXR dataset (Boecking et al., 2022), developed by Boecking et al., contains 1162 image-sentence pairs covering eight different cardiopulmonary radiological findings, with approximately equal samples for each finding. This dataset complements the MIMIC-CXR v.2 dataset, particularly in the areas of phrase grounding and radiological finding annotations. Specifically, it includes: reviewed and edited annotations, consisting of 1026 pairs of bounding boxes and corresponding phrases; and manually created annotations from scratch, consisting of 136 pairs of bounding boxes and corresponding phrases. Although the MS-CXR dataset (Boecking et al., 2022) lacks comprehensive radiology re-

ports, it still holds certain value in ARRG research. Specifically, the MS-CXR dataset (Boecking et al., 2022) can be utilized as a data source for radiology visual grounding and can also be employed to evaluate the performance of ARRG models in multitasking scenarios, such as Visual Question Answering (VQA) and Phrase Grounding (Bannur et al., 2024; Chen et al., 2024c; Li et al., 2023c; Park et al., 2024).

- **MS-CXR-T Dataset (Bannur et al., 2023a,b):** Bannur et al. extended and meticulously annotated the MIMIC-CXR v2 dataset to create MS-CXR-T (Bannur et al., 2023a,b). This multimodal benchmark dataset is designed to evaluate models in biomedical vision-language processing, specifically for temporal tasks in radiology, including temporal image classification and temporal sentence similarity analysis. The temporal image classification task comprises 1326 pairs of chest X-ray images, addressing five pathological findings: consolidation, edema, pleural effusion, pneumonia, and pneumothorax. Each pathology is labeled with one of three progression categories: "Improving, "Stable, or "Worsening. The temporal sentence similarity task includes 361 sentence pairs, annotated as either "Paraphrase or "Contradiction, capturing variations in temporal semantics. The goal of MS-CXR-T dataset (Bannur et al., 2023a,b) is to bridge the gap in benchmark datasets for temporal tasks in the biomedical domain, supporting the evaluation of vision and language models in quantifying disease progression and addressing time-sensitive challenges. The MS-CXR-T dataset (Bannur et al., 2023a,b) does not provide full radiology reports either.
- **RadGraph Dataset** (Jain et al., 2021): RadGraph (Jain et al., 2021) is a dataset focused on entity and relation extraction from radiology reports. It is based on reports from MIMIC-CXR dataset (Johnson et al., 2019a,b) and CheXpert dataset (Irvin et al., 2019), annotated by experienced radiologists following a carefully designed extraction framework. The dataset includes four entity types (anatomical structures and three observation categories) and three relation types (located_at, modify, and suggestive_of). It provides both manually annotated data and automatically generated annotations created using a deep learning model, covering entity and relation labels in 220,763 MIMIC-CXR (Johnson et al., 2019a,b) reports and 500 CheXpert dataset (Irvin et al., 2019) reports. Although the RadGraph dataset (Jain et al., 2021) itself does not directly include visual information such as X-ray images, the annotations can be mapped to corresponding images in the MIMIC-CXR (Johnson et al., 2019a,b) and CheXpert dataset (Irvin et al., 2019) datasets, enabling multimodal research that integrates textual and visual information.
- **RadGraph2 Dataset** (Dejl et al., 2024; Khanna et al., 2023): RadGraph2 (Dejl et al., 2024; Khanna et al., 2023) comprises 800 chest radiology reports annotated with fine-grained entities and relationships, building upon and enhancing the original RadGraph dataset (Jain et al., 2021). This dataset introduces new entity and relationship types to capture temporal changes, such as the progression, improvement, or complete resolution of medical conditions, as well as the placement, repositioning, or removal of medical devices. These additions address the limitation of previous datasets that focused solely on findings from single scans. The extracted information is represented as a knowledge graph, enabling structured and automated processing. In addition to these manually annotated reports, RadGraph2 dataset (Dejl et al., 2024; Khanna et al., 2023) includes over 220,000 automatically annotated reports generated by a high-performance benchmark model, which achieved F1 scores of 0.88 on the MIMIC-CXR-JPG dataset (Johnson et al., 2019a,b) and 0.74 on the CheXpert dataset (Irvin et al., 2019) dataset.

4.2.3. *Pathology detection in chest radiographs (PadChest) dataset (Bustos et al., 2020)*

PadChest dataset (Bustos et al., 2020) is a large-scale, high-resolution chest X-ray dataset containing over 160,000 images from 67,000 patients. These data were collected between 2009 and 2017 in the Radiology Department of San Juan Hospital in Spain and were interpreted and reported by physicians. The dataset includes 174 radiographic findings, 19 differential diagnoses, and 104 anatomical locations, with labels organized in a hierarchical classification system and mapped to the standardized Unified Medical Language System (UMLS) terminology. PadChest dataset (Bustos et al., 2020) is the first dataset to include radiology reports in Spanish and also provides bilingual support (Spanish and English), offering a unique opportunity for cross-lingual research, especially for validating model generalization in different language environments. To access the PadChest dataset (Bustos et al., 2020) need to fill out the application form and complete the approval process (BIMCV, 2025).

The PadChest dataset (Bustos et al., 2020) not only contains chest X-ray images but also provides corresponding fields, including: StudyDate, PatientSex, ViewPosition, Modality, Manufacturer, PhotometricInterpretation, PixelRepresentation, Data representation of the pixel samples, PixelAspectRatio, SpatialResolution, BitsStored, WindowCenter, WindowWidth, Rows, Columns, XRayTubeCurrent, ExposureTime, Duration of x-ray exposure, Exposure, ExposureInuAs, RelativeXRayExposure, and others. Fig. 12 is an example from the PadChest dataset (Bustos et al., 2020), where the top one is the chest X-ray image and the bottom one represents a subset of the corresponding fields. This dataset has been applied in several studies (Bannur et al., 2024; Castro et al., 2024).

4.2.4. *Digital database for screening mammography (DDSM) dataset (Heath et al., 2001)*

The Digital Database for Screening Mammography (DDSM) dataset (Heath et al., 2001) is a collaborative effort involving co-principal investigators from Massachusetts General Hospital (D. Kopans, R. Moore), the University of South Florida (K. Bowyer), and Sandia National Laboratories (P. Kegelmeyer). This dataset contains 2620 mammographic images of scanned films, stored in DICOM format, covering normal, benign, and malignant cases, and includes verified pathological information. The dataset also provides annotations of abnormal areas in the images by experts, such as the location and type of calcifications and masses, but does not include complete radiological reports. Access to this dataset requires a formal request. Although the DDSM dataset (Heath et al., 2001) does not include complete radiological diagnostic reports, it provides annotations of abnormal areas, which can be used to validate the model's performance in lesion detection tasks. These annotations allow the model to learn how to identify and localize lesion areas during the training process, thereby evaluating its accuracy and effectiveness in lesion detection (Kisilev et al., 2016).

4.2.5. *COVID-19 CT (China) dataset (Liu et al., 2021c)*

The COVID-19 CT (China) dataset (Liu et al., 2021c) is a CT dataset focused on COVID-19 cases. It contains 1104 chest CT images and 368 Chinese medical reports, sourced from the First Affiliated Hospital of Jinan University in Guangzhou and the Fifth Affiliated Hospital of Sun Yat-sen University in Zhuhai. The dataset includes 96 patients, with ages ranging from 10 months to 80 years. The reports provide detailed descriptions of pulmonary lesions, such as ground-glass opacities and consolidations, accompanied by relevant medical terminology tags. Each report is paired with multiple representative CT images to support the analysis and diagnosis of lesion areas. Access to the dataset requires signing the COVID-19 CT Dataset License Agreement and returning it to the dataset administrators. Fig. 13 is an example of the COVID-19 CT (China) dataset (Liu et al., 2021c). On the left are four lung CT images, and on the right is the corresponding medical image report in Chinese (with an English translation provided by the author). The medical terminologies in the "Findings section of the report are highlighted in red.

4.2.6. *CT-RATE dataset (Hamamci et al., 2024a)*

CT-RATE dataset (Hamamci et al., 2024a) is the first dataset to pair 3D medical images (chest) with their corresponding radiology reports. This dataset includes 25,692 non-contrast 3D chest CT scans from

Item	Description
ViewPosition	PA
Modality	DX
Report	cardiomegali . pinzamient ambos sen costofren sugest component derram pleural asoci . hili prominent probabl orig vascul . respect estudi previ fech 29 04 2016 con comp apreci infiltr alveol lsd nuev aparicion context pacient sugier orig cardiogen redistribucion aunqu sin pod descart orig inflamatori infecci fiebr leucocitosis correlacion con dat clinic .
Labels	['alveolar pattern', 'cardiomegaly', 'vascular hilar enlargement', 'pleural effusion', 'pneumonia', 'costophrenic angle blunting', 'vascular redistribution', 'heart insufficiency'
Localizations	['loc costophrenic angle', 'loc right upper lobe', 'loc hilar', 'loc cardiac', 'loc pleural']
labelCUIS	['C1332240' 'C0018800' 'C2073625' 'C0032285' 'C0742855' 'C0239041' 'C0018801']
LocalizationsCUIS	['C0230151' 'C1261074' 'C0205150' 'C1522601' 'C0032225']

Fig. 12. An example from the PadChest dataset (Bustos et al., 2020). The top image shows a chest X-ray, while the bottom content represents a subset of the corresponding fields.

21,304 unique patients and, through various reconstructions, expands to 50,188 volumes comprising over 14.3 million 2D slices. Each CT scan is accompanied by its associated radiology report, providing a robust foundation for research in medical imaging AI. It has been widely utilized in multiple studies, including (Chen et al., 2024a,b; Deng et al., 2024; Di Piazza, 2024; Hamamci et al., 2024c, 2025; Zhang et al., 2024b). However, it is important to clarify that while some articles cite (Hamamci et al., 2024b) when referring to this dataset, this citation is often misaligned. Upon further examination, the dataset is primarily based on the study referenced as (Hamamci et al., 2024a).

4.2.7. *Liver tumour segmentation (LiTS) dataset (Bilic et al., 2023)*

Liver Tumour Segmentation (LiTS) dataset (Bilic et al., 2023) is a 3D CT dataset focused on liver and liver tumour segmentation. The dataset was collected from seven different medical centers and includes 131 contrast-enhanced abdominal CT scans in the training set and 70 in the test set. This dataset does not include corresponding radiology reports. It is available for download online (Codalab Competitions, 2017), but access requires an application. In Tian et al.'s study, Chinese medical experts collaborated to create radiology reports in Chinese for the dataset (Tian et al., 2018). These reports described the liver and tumour in terms of shape, contour, and intensity, aiming to evaluate the impact of attention mechanisms on the performance of report generation.

4.3. *Private datasets*

Private datasets refer to datasets collected internally by hospitals, medical companies, research institutions, or enterprises but are not publicly available or authorized for external researchers.

4.3.1. *Breast cancer dataset (BCD2018) (Yang et al., 2021b)*

Although ultrasound is part of the field of medical imaging, it is traditionally not classified under radiology. However, considering that a major trend in ARRG research is the adoption of multimodal approaches-integrating various types of medical imaging and data, such as ultrasound, X-ray, and CT, to enhance diagnostic comprehensiveness and accuracy-it is therefore appropriate to also review datasets related to ultrasound in this context.

Breast Cancer Dataset (BCD2018) (Yang et al., 2021b) is a breast cancer ultrasound image dataset consisting of breast ultrasound images and their corresponding medical reports in Chinese. The data was collected from a professional medical institution, where experienced physicians carefully selected high-quality breast ultrasound images from the original data and wrote detailed medical reports. The dataset includes 5349 breast ultrasound images and 5349 Chinese medical reports, with each image paired with one corresponding report. The number of ultrasound images per patient ranges from 1 to 10. The author provided a link (Baidu , 2025) in the paper for accessing the dataset; however, the link is currently invalid. It may be possible to obtain access to the dataset by directly contacting the author. Therefore, in this review, we classify this dataset as private to reflect its inability to be accessed through public channels.

4.3.2. *BCD2019 Dataset (Wang et al., 2023a)*

BCD2019 dataset (Wang et al., 2023a) is a breast ultrasound dataset. With the assistance of medical experts, this dataset comprises 4384 carefully selected high-quality images and their corresponding 1363 Chinese medical reports. Each patient has between 1 and 10 images. To facilitate the extraction of domain knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004), these Chinese reports were pre-translated into English.

4.3.3. *BCT-CHR Dataset (Yang et al., 2021a)*

The BCT-CHR dataset (Yang et al., 2021a) is a brain CT dataset containing 2048 anonymized image-report pairs, with all corresponding reports written in Chinese. Each report consists of two sections: Findings and Impression, where the Findings section provides a detailed description of both normal and abnormal features observed in the images, while the Impression section summarizes the key medical conclusions.

4.3.4. *Gallbladder stone ultrasound (GS-Ultrasound) dataset (Zeng et al., 2024)*

The Gallbladder Stone Ultrasound (GS-Ultrasound) Dataset (Zeng et al., 2024) was introduced by Zeng et al. and was collected in collaboration with a hospital in Chongqing. This dataset comprises 6563

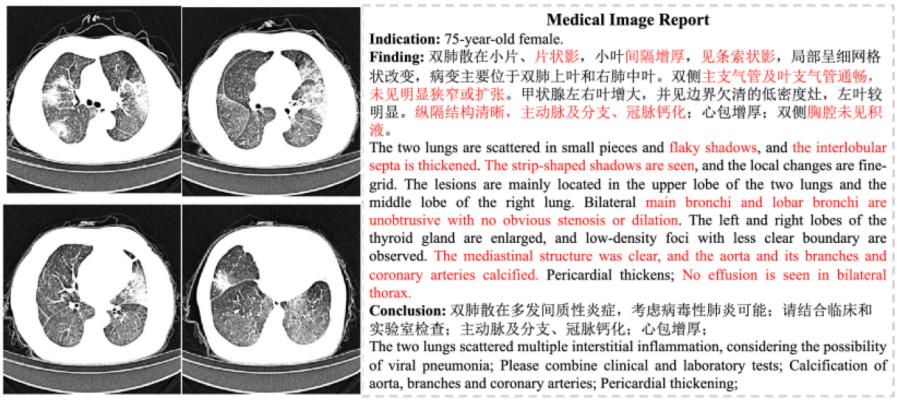


Fig. 13. It presents an example of the COVID-19 CT (China) dataset (Liu et al., 2021c). On the left are four lung CT images, and on the right is the corresponding medical image report in Chi- nese (with an English translation provided by the author). The medical terminologies in the "Findings section of the report are highlighted in red.

image-text pairs. Additionally, it includes 11 disease categories: gallbladder stone, liver cyst, hemangioma, normal liver, fatty liver, gallbladder polyp, normal gallbladder, kidney stone, renal cyst, hydronephrosis, and normal kidney.

4.3.5. *Fetal heart ultrasound (FH-Ultrasound) dataset (Zeng et al., 2024)*

The Fetal Heart Ultrasound (FH-Ultrasound) Dataset was proposed by Zeng et al. and was collected in collaboration with a hospital in Chongqing (Zeng et al., 2024). This dataset consists of 3300 image-text pairs. Additionally, it includes 11 fetal heart ultrasound views, namely abdominal cross section, four-chamber heart, left ventricular outflow tract, right ventricular outflow tract, three-vessel, three-vessel trachea, aortic arch, ductus arteriosus arch, cardiac aorta short axis, short axis of the ventricle, and superior and inferior vena cava.

4.3.6. *CX-CHR Dataset (Li et al., 2019a, 2018; Wang et al., 2020)*

The CX-CHR dataset (Li et al., 2019a, 2018; Wang et al., 2020) is a private chest X-ray dataset. Li et al. mentioned in article (Li et al., 2019a) that the CX-CHR dataset (Li et al., 2019a, 2018; Wang et al., 2020) consists of 35,609 patients and 45,598 images. Similarly, Wang et al. provided the same description in article (Wang et al., 2020). However, in another article (Li et al., 2018), Li et al. stated that the dataset includes 35,500 patients. Since CX-CHR (Li et al., 2019a, 2018; Wang et al., 2020) is a private dataset, the specific details cannot be directly verified. The difference in the number of patients may be attributed to updates made to the dataset at different points in time. The dataset contains corresponding radiology reports written in Chinese.

4.3.7. *JLiverCT dataset (Nishino et al., 2022)*

The JLiverCT dataset (Nishino et al., 2022) is a medical dataset focused on liver CT imaging, comprising 1083 Japanese reports that provide detailed descriptions of liver lesions and their appearances across different time phases, such as the arterial and portal venous phases during contrast-enhanced scans. Each report includes 65 lesion labels and corresponding time sequence information, making it a valuable resource for studying time-series diagnostic tasks and automated report generation.

These private datasets are valuable because they cover several ARRG scenarios that are less visible in public chest X-ray benchmarks, including breast ultrasound, fetal heart ultrasound, Chinese-language brain CT reporting, and liver CT reporting with phase information. Compared with standard chest X-ray report generation settings, these datasets reflect additional challenges such as multi-image-per-patient input, organ- and modality-specific terminology, Chinese-language reporting, and, in some CT tasks, the need to describe findings across different contrast-enhancement phases. At the same time, because they extend ARRG beyond a single imaging setting, they may also provide useful reference points for future modality-aware or broader multimodal ARRG research.

This section reviews 36 medical imaging datasets, including 16 public datasets, 13 restricted-access datasets, and 7 private datasets. Table 6 provides an overview of commonly used datasets for ARRG, but these resources do not occupy the same role in the pipeline. Some datasets can directly support end-to-end report generation because they provide paired medical images and complete radiology reports, whereas others are more appropriately regarded as auxiliary resources for pretraining, weak supervision, anatomical grounding, structured reporting, or evaluation. To make this distinction more explicit, Table 6 now labels the textual-information field according to whether a dataset is directly usable for ARRG (D), mainly auxiliary (A), or mainly evaluation- or annotation-oriented (E).

It is also important to distinguish the clinical role of pathology from its practical role in ARRG research. Although pathology may provide diagnostically important downstream confirmation and is often regarded as a clinical reference standard for many diseases, it does not usually function as the primary supervision target for ARRG in the same way as radiology report text. ARRG models are typically trained to generate radiology reports from imaging inputs, so paired radiology image-text datasets play the most direct role in report generation, whereas pathology data more often serves a complementary role for validation, correlation, or broader clinical interpretation.

Some of these datasets contain corresponding diagnostic reports, and the report languages are not limited to English. For example, the PadChest dataset (Bustos et al., 2020) includes Spanish reports; the CTRG-Brain-263K dataset (Tang et al., 2024), CTRG-Chest-548K dataset (Tang et al., 2024), and COV-CTR dataset (Li et al., 2023b) provide bilingual reports in Chinese and English; and the INbreast dataset (Moreira et al., 2012) contains Portuguese reports. Additionally, these datasets encompass a wide range of medical imaging modalities, including X-ray, mammography, CT, MRI, PET-CT, and ultrasound, providing essential resources for multimodal medical imaging research. In particular, datasets with multilingual diagnostic reports make it possible to examine model behaviour across different linguistic settings, while auxiliary datasets with labels, grounding annotations, or structured textual elements can support pretraining, supervision, and factual evaluation.

From a chronological perspective, the dataset landscape also reflects the evolution of the field, as summarized in Fig. 14. Early benchmarks such as IU X-ray (Demner-Fushman et al., 2016) made paired chest X-ray report generation possible; later large-scale restricted resources such as MIMIC-CXR and its derivatives enabled broader benchmarking and

Table 6
An overview of commonly used datasets for ARRG, including public, restricted-access, and private datasets. “U-R”, “S-R”, “S-R-A”, and “A-R” represent “unstructured reports”, “structured reports”, “structured report annotation”, and “annotated reports”, respectively. “D”, “A”, and “E” denote datasets that can be used directly for ARRG, datasets mainly used as auxiliary resources (e.g., pretraining, weak supervision, or grounding), and datasets mainly used for evaluation or annotation purposes, respectively. “EN, PT, ZH, ES, JA” denote “English, Portuguese, Chinese, Spanish, and Japanese,” respectively.

Dataset Name	Access	2D/ 3D	Image Modalities	Image Count	Textual Information	Text count	Text Language
IU X-ray (Demner-Fushman et al., 2016)	Public	2D	X-ray (Chest)	7470	U-R (D)	3955	EN
INbreast (Moreira et al., 2012)	Public	2D	Mammography	410	U-R (D)	115	PT
COV-CTR (Li et al., 2023b)	Public	2D	CT (Chest)	728	U-R (D)	728	ZH & EN
CTRG-Brain-263K (Tang et al., 2024)	Public	2D	CT (Brain)	263,670	U-R (D)	10,009	ZH & EN
CTRG-Chest-548K (Tang et al., 2024)	Public	2D	CT (Chest)	548,696	U-R (D)	1804	ZH & EN
Rad-ReStruct (Pellegrini et al., 2023)	Public	2D	X-ray (Chest)	3720	S-R-A (A)	3597	EN
Shenzhen-set (Candemir et al., 2013) (Jaeger et al., 2013)	Public	2D	X-ray (Chest)	662	Annotations (A)	662	EN
PEIR (PEIR Digital Library, 2025)	Public	2D	Multimodal	4732	Sentences (A)	4732	EN
ROCO (Pelka et al., 2018)	Public	2D	Multimodal	81,825	Sentences (A)	81,825	EN
ROCOv2 (Rückert et al., 2024)	Public	2D	Multimodal	79,789	Sentences (A)	79,789	EN
CheXpert (Irvin et al., 2019)	Public	2D	X-ray (Chest)	224,316	Labels (A)	14	EN
CheXpert Plus (Chambon et al., 2024), (Medicine, 2024)	Public	2D	X-ray (Chest)	223,462	U-R (D)	187,711	EN
ChestX-ray8 & ChestX-ray14 (Wang et al., 2017)	Public	2D	X-ray (Chest)	108,948	Labels (A)	8/ 14	EN
DeepLesion (Yan et al., 2018)	Public	2D	CT (Whole Body)	32,120	Annotations (A)	32,735	EN
Lumbar Spine MRI (Al-Kafri et al., 2019; Sudirman et al., 2019a)	Public	3D	MRI (Lumbar Spine)	575	U-R (D)	575	EN
ImageCLEF2015 Liver CT Annotation (Marvasti et al., 2015)	Public	3D	CT (Liver)	60	S-R (D)	60	EN
MIMIC-CXR/ MIMIC-CXR-JPG (Johnson et al., 2019a,b)	Restricted	2D	X-ray (Chest)	377,110	U-R (D)	227,835	EN
MIMIC-ABN (Ni et al., 2020)	Restricted	2D	X-ray (Chest)	38,551	U-R (D)	38,551	EN
CXR-PRO (Ramesh et al., 2022a,b)	Restricted	2D	X-ray (Chest)	374,139	U-R (D)	374,139	EN
RadGraph (Jain et al., 2021)	Restricted	2D	None	None	Annotations (E)	Unknown	EN
RadGraph2 (Dejl et al., 2024; Khanna et al., 2023)	Restricted	2D	None	None	A-R (E)	220,800	EN
PadChest (Bustos et al., 2020)	Restricted	2D	X-ray (Chest)	160,868	U-R (D)	109,931	ES & EN
COVID-19 CT (China) (Liu et al., 2021c)	Restricted	2D	CT (Chest)	1,104	U-R (D)	368	ZH
Chest ImaGenome (Krishna et al., 2017)	Restricted	2D	X-ray (Chest)	242,072	Annotations (A)	Unknown	EN
MS-CXR (Boecking et al., 2022)	Restricted	2D	X-ray (Chest)	1162	Sentences (E)	1162	EN
MS-CXR-T (Bannur et al., 2023a,b)	Restricted	2D	X-ray (Chest)	1326	Annotations (E)	1326	EN
DDSM (Heath et al., 2001)	Restricted	2D	Mammography	10,480	Annotations (A)	Unknown	EN
CT-RATE (Hamamci et al., 2024a)	Restricted	3D	CT (Chest)	25,692	U-R (D)	21,304	EN
LiTS (Bilic et al., 2023)	Restricted	3D	CT (Liver)	201	None (A)	None	None
BCD2018 (Yang et al., 2021b)	Private	2D	US	5349	U-R (D)	5349	ZH
BCD2019 (Wang et al., 2023a)	Private	2D	US	4384	U-R (D)	1363	ZH & EN
BCT-CHR (Yang et al., 2021a)	Private	2D	CT (Brain)	2048	U-R (D)	2048	ZH
GS-Ultrasound (Zeng et al., 2024)	Private	2D	Ultrasound	6563	U-R (D)	6563	ZH
FH-Ultrasound (Zeng et al., 2024)	Private	2D	Ultrasound	3300	U-R (D)	3300	ZH
CX-CHR (Li et al., 2019a, 2018; Wang et al., 2020)	Private	2D	X-ray (Chest)	45,598	U-R (D)	45,598	ZH
JLiverCT (Nishino et al., 2022)	Private	3D	CT (Liver)	1083	U-R (D)	1083	JA

auxiliary supervision; and more recent datasets such as CheXpert Plus, CTRG-Brain-263K, CTRG-Chest-548K, and CT-RATE indicate a shift toward larger-scale, multilingual, and increasingly 3D or clinically richer resources. This trajectory helps explain why recent ARRG systems place greater emphasis on multimodality, grounding, and factual control than earlier chest X-ray captioning-style models.

However, the datasets used in ARRG research still exhibit several limitations:

1. Complex data access procedures
   Accessing medical imaging datasets often involves lengthy approval processes, limiting widespread use by researchers. For instance, obtaining access to the MIMIC-CXR dataset (Johnson et al., 2019a,b) and MIMIC Derivatives dataset requires submitting a detailed research proposal and awaiting approval. Similar restrictions apply to the DDSM dataset (Heath et al., 2001), despite its extensive use in mammographic imaging research. While these strict ethical review procedures help protect patient privacy and ensure compliance with data usage regulations, they also reduce data accessibility, thereby hindering innovation in the field.
2. Insufficient data diversity
   Most publicly available datasets, such as IU X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019a,b), primarily consist of 2D chest X-rays. These datasets have inherent limitations, including low resolution, making it difficult to capture fine-grained pathological details. Overlapping anatomical structures, which obscure clear boundaries between tissues and lesions, negatively impacting visualization and diagnostic accuracy.

   Although 3D imaging provides richer anatomical information (e.g., chest CT scans can detect lung abnormalities, mediastinal lymph node changes, and bone lesions), the number of publicly available 3D datasets remains limited. For example, the COV-CTR dataset (Li et al., 2023b) contains only 728 COVID-19 chest CT scans, making it insufficient for large-scale training. The Lumbar Spine MRI dataset (Al-Kafri et al., 2019; Sudirman et al., 2019a) includes only 575 multi-slice lumbar spine MRI scans, restricting its applicability. Despite these limitations, 3D imaging plays a crucial role in clinical applications, such as detecting rib fractures or bone metastases in cancer patients, tasks that are often challenging to accomplish using 2D imaging alone. Additionally, the complexity of 3D imaging increases the workload of radiologists, as generating reports for 3D

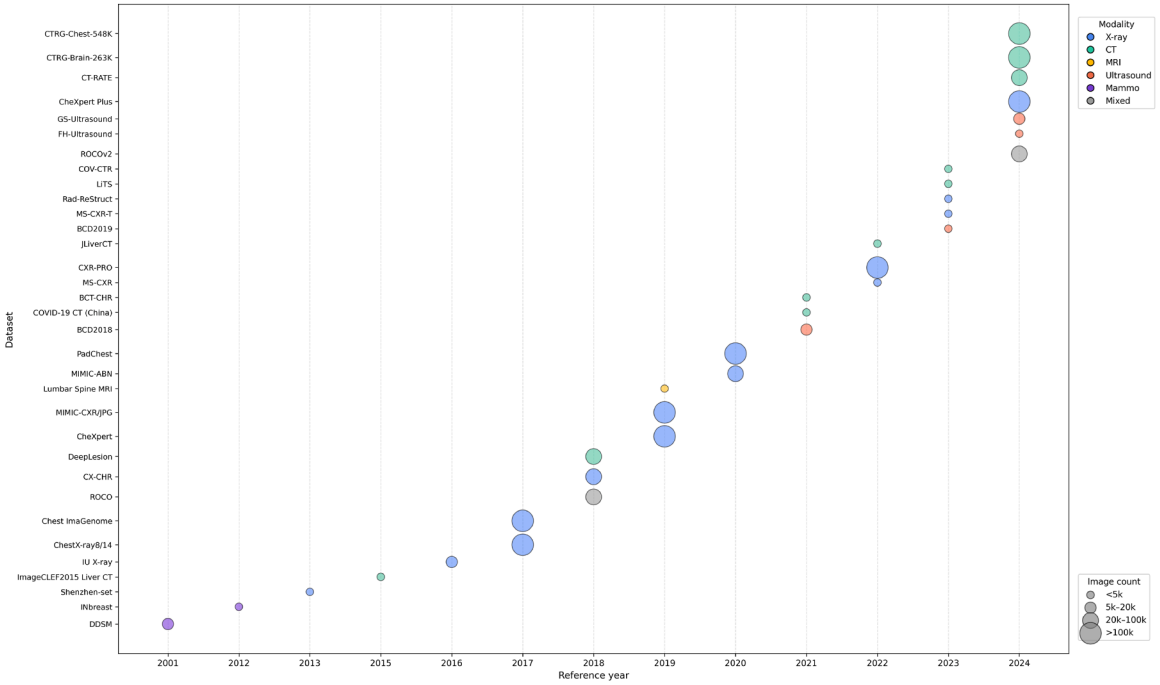


Fig. 14. Chronological distribution of the datasets reviewed in this article. Bubble size reflects approximate image count, and colour indicates imaging modality. The year shown for each dataset follows the publication year of the corresponding cited reference in this review. Datasets for which a publication year or image-related information could not be reliably determined from the available cited source were not included in the figure. A full overview of all datasets is provided in Table 6.

scans is more time-consuming. Therefore, developing ARRG systems capable of generating high-quality 3D imaging reports is both highly valuable and urgently needed.

3. Data distribution bias
   Publicly available medical imaging datasets suffer from imbalanced disease distributions, which primarily manifest in the following ways:
   - Limited representation of rare diseases
     For example, IU X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019a,b) mainly feature common diseases such as pneumonia and pleural effusion, whereas rare diseases like interstitial lung disease are underrepresented. This imbalance can lead to poor generalization of models when handling rare or complex cases, reducing their clinical applicability.
   - Lack of datasets for other organs
     Most publicly available datasets remain focused on chest imaging (e.g., X-ray and CT), while large-scale datasets for other anatomical regions, such as the brain, cardiovascular system, abdomen, and musculoskeletal system, remain scarce. In recent years, progress has been made in addressing data distribution bias. For instance, Tang et al. introduced the CTRG-Brain-263K (Tang et al., 2024) dataset, a large-scale, publicly available brain CT dataset that includes bilingual Chinese-English reports and maintains a 1.5:1 ratio of normal to abnormal cases. This dataset helps mitigate data imbalance to some extent. However, further expansion of disease categories in medical imaging datasets is still required to enhance model robustness and generalization.
4. Small dataset sizes
   Many datasets used in ARRG research remain relatively small, leading to several challenges. For example, limited generalization ability, making it difficult for models to adapt to various clinical settings. Increased risk of overfitting, resulting in poor performance on unseen data. Although datasets like MIMIC-CXR dataset (Johnson et al., 2019a,b) and CheXpert dataset (Irvin et al., 2019) provide large-scale X-ray images and reports, many ARRG studies still rely on smaller datasets. For example, the IU X-ray dataset (Demner-Fushman et al., 2016) contains only 7470 reports. Additionally, even large-scale datasets often suffer from inconsistent labeling quality, as some rely on automatically extracted labels, which may introduce noisy data that negatively impact model performance.

Therefore, improving annotation quality for ARRG tasks and expanding large-scale, multimodal medical imaging datasets remain critical directions for future research.

5. ARRG Methodological paradigms and enhancement mechanisms

ARRG can be regarded as a specialized extension of image captioning (Liao et al., 2023; Luan et al., 2023; Sloan et al., 2024; Wang et al., 2022a), aimed at generating diagnostic reports from medical images (e.g., X-ray, CT, MRI, US). Unlike conventional image captioning, which primarily describes generic visual elements, ARRG requires the integration of computer vision (CV) (Elyan et al., 2022; Javaid et al., 2024) and natural language processing (NLP) (Chng et al., 2023; Zhang et al., 2019b), along with domain-specific medical knowledge, to generate clinically relevant reports with appropriate factual content, ordering, and level of detail.

Given its strong connection to image captioning, ARRG research methodologies have drawn inspiration from captioning techniques while adapting them to the specific challenges of medical imaging and reporting. In this section, we distinguish between methodological paradigms and enhancement mechanisms. The former includes template-based, retrieval-based, encoder-decoder, foundation-model-based, and hybrid approaches. Within the encoder-decoder lineage, CNNs are mainly used as visual encoders and RNNs or hierarchical RNNs as sequence decoders. The latter includes attention, multimodal fusion, reinforcement learning, and knowledge integration, each of which may function either as a central architectural component or as an auxiliary depending on the model family.

5.1. Methodological paradigms

5.1.1. Template-based models

The core idea of the template-based methods lies in its generation mechanism based on predefined templates. Essentially, templates are pre-designed sentence or paragraph structures that include a set of slots to be filled with content. By extracting information from the input data and populating these slots, it becomes possible to generate grammatically correct and semantically coherent text.

Pino et al. proposed a model named CNN-TRG, which generates reports by detecting abnormalities in images and relying on predefined templates (Pino et al., 2021). The report generation process involves using a CNN for multi-label classification to detect the presence or absence of 13 abnormalities. Based on the classification results, corresponding descriptions (e.g., the presence or absence of abnormalities) are selected from the template pool. Finally, these sentences are combined to create a complete medical report. Abela et al. proposed a template-based report generation system called Transformer-Template Report Generation (T-TRG) and conducted comparative experiments with the CNN-TRG model, based on convolutional neural networks, and a baseline Encoder-Decoder model (Abela et al., 2022). Both the T-TRG and CNN-TRG models utilize multi-label classifiers to detect abnormalities in chest X-rays and employ a template subsystem to retrieve matching descriptions from predefined sentence templates. The retrieved sentences are then concatenated to generate the final report.

Template-based methods rely heavily on predefined templates, which limits their flexibility and creativity in generating diverse or novel content (Messina et al., 2022). They require significant manual effort to design (Messina et al., 2022), optimize, and maintain the templates, making them time-consuming and labor-intensive. These methods struggle to handle complex or atypical cases (Beddiar et al., 2023), as their outputs are constrained by the templates' coverage and structure.

5.1.2. Retrieval-based models

Retrieval-based methods were commonly used in the early stages of image captioning. The core idea is to retrieve the most relevant sentence or description from a pre-established database (corpus) based on the similarity between the input image features and the features stored in the database (Reale-Nosei et al., 2024). Similarly to image captioning, the retrieval-based methods have also been applied in ARRG research. The methods do not have a fixed architecture (Liao et al., 2023). The most critical challenge lies in designing effective retrieval strategies to extract image features and match them with corresponding sentence templates.

Zhang et al. proposed a transfer learning approach that combines multi-label classification with retrieval methods (Zhang et al., 2018). Their method retrieves visually similar images based on colour and texture features, selects the descriptions of the top three most similar images, and combines these descriptions to generate a new report. Ni et al. proposed a method that aligns visual and semantic features (Ni et al., 2020). By incorporating a weighted attention mechanism, it calculates visual-semantic similarity using the squared l2-normalized Euclidean distance. Kougia V et al. adopted a different approach by calculating the cosine similarity between the visual embeddings of the input image and sentences, selecting the most relevant sentence (Kougia et al., 2021). Syeda-Mahmood et al. introduced a domain-aware retrieval method (Syeda-Mahmood et al., 2020). This method first learns fine-grained features of lesions in medical images to describe the image content, and then uses these extracted features to retrieve the most similar reports from a large database, followed by customization and optimization. In the study by Charalampakos et al., a retrieval method based on k-nearest neighbors (k-NN) was introduced (Charalampakos et al., 2021). The core idea is to compute the cosine similarity between the input image's embedding and the embeddings of images in the training set to match and retrieve the most similar image. Yang et al. proposed a hierarchical retrieval mechanism called MedWriter, designed to extract both report-level and sentence-level templates during the process of medical report generation (Yang et al., 2021c). The mechanism integrates three key components. First, the Visual-Language Retrieval (VLR) module operates at the report level, leveraging the visual features of the input image to retrieve the most semantically relevant report templates from a retrieval pool. Next, the Language-Language Retrieval (LLR) module functions at the sentence level, identifying candidate sentences from the retrieval pool that are most likely to serve as the next sentence, thereby ensuring logical coherence and consistency in the report. Finally, the Hierarchical LSTM Decoder combines the features generated by the VLR and LLR modules along with the image features to generate the medical report sentence by sentence, producing accurate and clinically relevant outputs.

Retrieval-based methods primarily rely on pre-established large-scale databases, generating outputs by matching input data with existing content in the database based on similarity. However, these methods have several drawbacks: they lack flexibility and creativity, making it impossible to generate new content beyond the scope of the database (Beddiar et al., 2023; Liu et al., 2025); they are heavily dependent on the quality and coverage of the database, struggling to handle unseen data or rare cases (Beddiar et al., 2023).

5.1.3. Encoder-decoder paradigms

CNNs primarily act as visual encoders that transform medical images into feature representations, while RNNs, LSTMs, GRUs, and hierarchical RNN variants mainly serve as sequence decoders that organize these representations into report text. In the early stages of ARRG research, CNNs were widely used to extract visual features from medical images. ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) are among the most representative models in this context. ResNet (He et al., 2016) introduces the concept of residual connections (skip connections), which form the foundation of the residual learning framework. This innovation significantly alleviates the vanishing gradient problem and the degradation issue in deep neural network training, enabling the development of deeper and more expressive networks. Specifically, ResNet (He et al., 2016) utilizes residual connections to allow the network to learn residual mappings rather than directly learning the full transformation. This approach reduces the learning difficulty and facilitates the learning of identity mappings, effectively addressing the degradation problem in deep networks. DenseNet (Huang et al., 2017), on the other hand, is composed of multiple dense blocks, where each layer within a dense block adopts dense connectivity-meaning that the input to each layer includes the outputs of all preceding layers. This design enhances feature propagation, allowing low-level features to be directly reused by higher layers; mitigates the vanishing gradient problem, improving gradient flow throughout the network; and enhances parameter efficiency, encouraging the network to learn more effective feature representations. Additionally, the dense connectivity inherently promotes feature reuse.

In ARRG research, for example, in the experiment conducted by Yin et al., a DenseNet variant with GLP (global label pooling) was used for feature extraction (Yin et al., 2019). Specifically, DenseNet was employed to extract the feature map, followed by the GLP (Global Label Pooling) mechanism to generate a label heat map. Finally, global max pooling was applied to obtain the prediction probability for each label. Yuan et al. proposed a typical encoder-decoder-based ARRG model (Yuan et al., 2019), in which the visual feature extraction component employs a ResNet-152 model (He et al., 2016) pre-trained on the CheXpert dataset (Irvin et al., 2019). In Liu et al.'s study, ResNet-152 was utilized for medical image feature extraction in the visual extraction phase, while the Posterior Knowledge Explorer (PoKE) was employed to further refine and highlight abnormal regions (Liu et al., 2021a). Specifically, PoKE integrates disease topic tags (Topic Bag) and multi-head attention (MHA) to enable the model to focus on abnormal areas in medical images, rather than distributing attention evenly across the entire image, as traditional methods do. In the AlignTransformer model proposed by You et al., the Align Hierarchical Attention (AHA) module

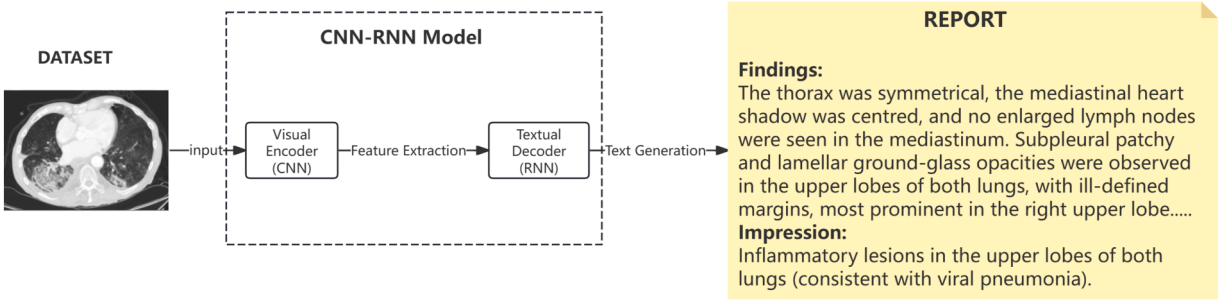


Fig. 15. A schematic representation of the ARRG framework based on the CNN-RNN architecture (Pang et al., 2023).

aligns the visual features extracted by ResNet-50 with the disease tags predicted by the multi-label classification network at different granular levels, thereby generating disease-related visual features (You et al., 2021). Li et al. proposed a model called ASGK (Auxiliary Signal-Guided Knowledge Encoder-Decoder) (Li et al., 2023b). In their study, they employed DenseNet-121 for feature extraction, obtaining visual features from medical images.

Although CNNs played a crucial role in early ARRG research, they primarily focus on local features and have limited ability to capture long-range global contextual information (Liao et al., 2023). This helps explain why later encoder-decoder systems increasingly incorporated hierarchical decoders, attention mechanisms, and Transformer-based components.

Early encoder-decoder systems often employed RNN as decoders. In early ARRG research, they were the dominant sequence generators because they could model report generation step by step and preserve short-range contextual dependencies. With the rapid advancement of deep learning technology, CNN-RNN encoder-decoder systems were widely adopted in image captioning tasks (Vinyals et al., 2015; Xu, 2015; Yao et al., 2017). Similarly, in ARRG, numerous studies employed CNN-recurrent encoder-decoder pipelines (Shin et al., 2016b; Sun et al., 2019; Wang et al., 2018; Yin et al., 2019; Yuan et al., 2019; Zeng et al., 2020b). In this architecture (Fig. 15), CNNs are responsible for extracting visual features from images, while RNNs and their variants (e.g., LSTM Hochreiter, 1997 and GRU Cho et al., 2014; Chung et al., 2014, 2015) process sequential information, ensuring that the generated text effectively incorporates contextual information to enhance the coherence and readability of reports. Related CNN-RNN designs were also explored in adjacent medical image captioning or interpretation settings, including ultrasound-based studies and visual interpretation work (Alsharid et al., 2019; Li et al., 2019b; Rodin et al., 2019; Zeng et al., 2020a).

In one of the earliest ARRG studies, Shin et al. proposed a method based on the CNN-RNN architecture for the classification and annotation of chest X-rays (Shin et al., 2016b). The system employed two CNN models, Network-in-Network (NIN) (Lin, 2013) and GoogLeNet (Szegedy et al., 2015), as encoders to classify the images into 17 categories, including normal and specific disease classes. Additionally, through the introduction of a Recurrent Cascade Model, the system refined the disease categories into 57 subcategories by integrating both image and textual context. RNNs, including LSTM (Hochreiter, 1997) and GRU (Cho et al., 2014; Chung et al., 2014, 2015), were then utilized to generate text descriptions associated with the identified diseases. Xue et al. proposed a multimodal recurrent model (Xue et al., 2018). In the text generation component of this model, a recursive sentence generation strategy is adopted. Specifically, an LSTM decoder is first used to generate the initial sentence based on the global features of the medical image. Subsequently, each new sentence is generated by incorporating the semantic features of the preceding sentence, which are encoded using a Bidirectional Long Short-Term Memory (Bi-LSTM) network, along with the local features of the image. Furthermore, some studies (Harzig et al., 2019; Huang et al., 2019; Jing et al., 2017; Krause et al., 2017; Yin et al., 2019; Yuan et al., 2019; Zhang et al., 2020) have improved upon RNNs by introducing hierarchical recurrent neural networks (HRNNs). Compared to standard RNNs, HRNNs adopt a hierarchical modeling strategy that enables more effective capture of the structural characteristics of long texts. Specifically, HRNNs employ a multi-layer RNN architecture to model the hierarchical structure of text: first, a sentence-level decoder generates a sentence-level semantic representation (i.e., a topic vector); subsequently, a word-level decoder takes this topic vector as input and sequentially generates individual words. Since radiology reports are typically organized hierarchically into paragraphs, sentences, and words, HRNNs can better capture linguistic features in reports, resulting in more coherent and semantically consistent text generation.

Jing et al. proposed a multi-task learning framework that includes two tasks: label prediction and paragraph generation (Jing et al., 2017). In their experiments, they utilized a Hierarchical LSTM model, which consists of two levels: sentence-level and word-level structures, to generate long-paragraph report texts. Yin et al. also employed an HRNN for text generation (Yin et al., 2019). The model consists of a sentence RNN, responsible for generating topic vectors, and a word RNN, which generates specific sentences based on these topic vectors. In the sentence RNN, a topic matching mechanism is utilized to enable the model to learn how to map topic vectors and ground truth sentences into the same semantic space. This mechanism ensures that the topic vectors generated by the sentence RNN are more semantically aligned with the ground truth sentences. In Huang et al.'s ARRG research, the text generation component also employs a Hierarchical LSTM (HLSTM) approach (Huang et al., 2019). This model consists of a sentence-level LSTM and a word-level LSTM, designed with a hierarchical decoder to effectively handle the complexity of long-text generation tasks. The sentence-level LSTM generates a topic vector for each sentence, ensuring the coherence of paragraph structure, while the word-level LSTM generates specific words based on the topic vector to construct complete sentences. This hierarchical structure not only overcomes the gradient vanishing problem commonly observed in traditional single-layer LSTM models for long-text generation but also significantly enhances the semantic consistency and structural coherence of the generated text, providing a more robust solution for complex paragraph generation tasks. Many studies have employed similar methods in text generation, such as Yuan et al. (2019) and Zhang et al. (2020), typically utilizing a two-layer LSTM architecture, consisting of a topic-level (or sentence-level) LSTM and a word-level LSTM. The topic-level LSTM predicts the theme of each sentence, while the sentence-level LSTM generates a topic vector for each sentence. Then, the word-level LSTM generates the specific sentence based on the predicted theme of each sentence or topic vector.

To address the issue of data imbalance in medical report generation and enhance sentence diversity, especially for more accurate generation of abnormal sentences, Harzig et al. proposed a Dual Word LSTM model (Harzig et al., 2019). This model adopts a dual-word LSTM structure, comprising an abnormal-word LSTM and a normal-word LSTM. Two independent word LSTMs are trained separately for normal and abnormal features to generate normal and abnormal descriptions, respec-

tively. These individually generated sentences are then integrated into a complete report. This design not only effectively mitigates the data imbalance problem but also significantly improves sentence diversity and the accuracy of abnormal case descriptions. Most studies adopt RNNs as decoders, such as LSTM (Hochreiter, 1997) and GRU (Cho et al., 2014; Chung et al., 2014, 2015). These models are often tailored to specific tasks to better meet the requirements of diagnostic report generation. Research has shown that RNNs, particularly LSTMs, perform exceptionally well in generating single-sentence descriptions (Reale-Nosei et al., 2024). It still has certain limitations. For example, standard RNNs often suffer from gradient vanishing or explosion when modeling long sequences, making it difficult to effectively capture long-range dependencies (You et al., 2021). While LSTMs (Hochreiter, 1997) and GRUs (Cho et al., 2014; Chung et al., 2014, 2015) mitigate the vanishing gradient problem through gating mechanisms, they still face sequential computation constraints, making it difficult to efficiently process long texts in parallel. Furthermore, their limited ability to retain long-term contextual information restricts their performance when generating paragraph-level text, leading to challenges in maintaining global coherence (Kaur et al., 2022).

HRNN, such as HLSTM, alleviate the aforementioned issues to some extent but do not completely resolve them. By adopting a hierarchical structure, text generation is divided into sentence-level and word-level processes, enabling the model to more effectively capture global semantic information and reducing the sequence length that a single-layer RNN needs to process, thereby improving computational efficiency to a certain degree. Moreover, this hierarchical modeling approach helps maintain cross-sentence contextual consistency, enhancing text coherence and structured representation, particularly outperforming standard LSTMs in paragraph-level text generation. However, despite mitigating information loss through hierarchical decoding mechanisms, HRNN still relies on the sequential nature of RNN computation. As a result, it remains susceptible to the vanishing gradient problem when handling long texts, making it difficult to fully retain long-range dependencies. Additionally, unlike Transformer models, HRNN cannot leverage parallel computation efficiently, which limits its computational efficiency when processing extremely long texts. Overall, HRNN and HLSTM offer advantages over standard LSTM in tasks such as radiology report generation, but for tasks requiring long-range dependency modeling and higher computational efficiency, Transformer remains the superior choice.

5.1.4. Foundation models

Although foundation-model-based approaches are discussed here as a distinct contemporary paradigm, most of them still follow an encoder-decoder logic in architectural terms. What distinguishes them from earlier encoder-decoder systems is not the complete abandonment of that structure, but the integration of large-scale pretraining, Transformer-based backbones, and stronger cross-modal representation learning. (Alfarghaly et al., 2021; Li et al., 2024b; Nicolson et al., 2023; Nooralahzadeh et al., 2021; Zhou et al., 2022) Foundation models, characterized by their large-scale pretraining and adaptability to diverse downstream tasks, have revolutionized the field of natural language processing (NLP) and computer vision (CV). These models, pretrained on extensive datasets, can be fine-tuned on smaller domain-specific datasets, effectively addressing the challenge of data scarcity in specialized fields such as medical imaging. Among these foundation models, the Transformer architecture (Vaswani, 2017) has emerged as a cornerstone due to its versatility and scalability. Its success in NLP tasks has inspired its adaptation to vision tasks, leading to the development of models like Vision Transformer (ViT) (Dosovitskiy, 2020). In the context of ARRG, foundation models (including Transformer-based architectures, BERT Kenton & Toutanova, 2019, GPT-3 Brown et al., 2020) have been applied in two primary directions: integrating with vision models to enable cross-modal conversion from medical images to text, and fine-tuning on medical datasets (e.g., MIMIC-CXR Johnson et al., 2019a,b) to enhance specialization and accuracy.

The Transformer (Vaswani, 2017) was originally designed for natural language processing tasks. On the one hand, due to its powerful sequence modeling capabilities and attention mechanism, it has also been widely applied to image feature extraction. In the vision domain, the Vision Transformer (ViT) model (Dosovitskiy, 2020) divides an image into small patches and then converts these patches into a sequence to be input into the model. The core of the Transformer is the self-attention mechanism, which allows the model to not only focus on local information but also capture the global relationships between different regions of the image. The Transformer is capable of effectively capturing long-range dependencies between different areas of the image, thus better understanding the overall structure and semantic information of the image. Transformer has achieved significant results in various computer vision tasks such as image classification, object detection, and semantic segmentation. As a result, an increasing number of researchers are exploring the application of Transformer in image feature extraction. Nooralahzadeh et al. proposed a progressive transformer-based model, in which visual feature extraction is performed by DenseNet to capture local structural features and convert them into a set of high-dimensional feature vectors (Nooralahzadeh et al., 2021). These feature vectors are then fed into ViLM (Meshed-Memory Transformer,M2 Transformer) to model the relationship between visual features and linguistic descriptions and further extract global concepts. Wang et al. proposed the ME-Transformer model, which introduces a multi-expert mechanism to enhance radiology report generation (Wang et al., 2023b). Its core architecture includes an expert encoder and an expert decoder: the expert encoder integrates a Vision Transformer (ViT) with a bilinear attention module, enabling learnable expert tokens to interact with visual tokens through high-order interactions, capturing fine-grained features from different regions of medical images, while orthogonal loss encourages the expert tokens to learn complementary information. Wu et al. proposed a generalist foundation model for radiology, named RadFM, which is capable of processing both 2D and 3D medical imaging data (Wu et al., 2023b). The model's visual encoder adopts a 3D Vision Transformer (3D ViT), which is designed for feature extraction from both 2D and 3D medical images. For 2D images (e.g., X-rays), the model expands them into a 3D format to align with the input requirements of the 3D ViT.

On the other hand, Transformers (Vaswani, 2017) leverage self-attention mechanisms, enabling efficient parallel computation. In addition to faster training speeds, they fully utilize the parallel computing power of GPUs. Their remarkable success in image captioning tasks (Herdade et al., 2019; Yu et al., 2019) has led to increasing adoption in ARRG, where they have gradually become a research hotspot. In the study by Lovelace et al., a pre-trained DenseNet-121 was used to extract features from chest X-ray images, combined with a Transformer encoder and decoder to achieve end-to-end radiology report generation (Lovelace & Mortazavi, 2020).

A differentiable CheXpert dataset (Irvin et al., 2019) label extraction mechanism and clinical coherence optimization strategy were introduced to generate semantically consistent clinical reports. Chen et al. integrated memory into the Transformer decoder to record historical information during the generation process and effectively leverage this information through multi-head attention, thereby enhancing the decoder's performance in long-text generation tasks (Chen et al., 2020). In the study by Yan et al. (2021), weakly supervised contrastive (WCL) learning, based on the memory-driven transformer proposed by Chen et al. (2020) to enhance the capability of ARRG models, was introduced. Specifically, WCL employs ChexBERT (Smit et al., 2020) and K-Means for report clustering, achieving weakly supervised semantic grouping to better select "hard negative samples" in contrastive learning. ChexBERT (Smit et al., 2020), a BERT variant tailored for medical text, is used to extract text embeddings from reports, while K-Means clustering groups reports based on semantic similarity. This allows WCL to distinguish between "easy negative samples" (clearly incorrect reports) and "hard negative samples" (semantically similar but incorrect

reports). Alfarghaly et al. proposed a model called CDGPT2, which uses a pre-trained distilGPT2 (Sanh, 2019) decoder conditioned on visual features (extracted by ChexNet) and semantic features (derived from label embeddings) to generate complete reports (Alfarghaly et al., 2021). Nooralahzadeh et al. proposed a Transformer-based progressive generation model that divides the radiology report generation task into two stages (Nooralahzadeh et al., 2021): in the first stage, a Visual-Language Model (ViLM) extracts high-level concepts from chest X-ray images using the Meshed-Memory Transformer (M2 Transformer) with memory-augmented encoders and decoders to generate global information; in the second stage, a pre-trained BART language model further refines these high-level concepts into a complete and coherent radiology report, achieving progressive generation while ensuring semantic consistency and clinical relevance. Zhou et al. proposed a cross-supervised learning model called REFERS, which uses a Radiograph Transformer to process image features, fuses multi-view images from patient studies to generate a unified representation, and employs a Report Transformer to generate corresponding radiology reports, while contrastive loss is applied to enhance the consistency between image and text representations (Zhou et al., 2022). Chen et al.'s experiment adopted a Transformer-based encoder-decoder architecture for text generation (Chen et al., 2022). In this stage, they also incorporated a cross-modal memory network (CMN), which enhances the cross-modal understanding capability of the Transformer decoder through the mechanisms of memory querying and memory responding. In the decoder of the METransformer model proposed by Wang et al., expert tokens guide the generation of multiple corresponding reports (Wang et al., 2023b). Each expert token interacts with both visual tokens and word tokens to generate a candidate report. In other words, if the model has M expert tokens, it will generate M different candidate reports.

Other foundation models have also made significant contributions to ARRG. These models, pretrained on vast corpora of text data, bring semantic understanding and generative capabilities to the task of medical report generation. For example, In Nicolson et al.'s experiment, the pre-trained language model distilGPT2 (Sanh, 2019) was utilized to initialize the decoder using warm starting techniques, enabling it to inherit the semantic understanding and generative capabilities of a general language model (Nicolson et al., 2023). The model was then fine-tuned on medical datasets such as MIMIC-CXR (Johnson et al., 2019a,b) and IU X-ray dataset (Demner-Fushman et al., 2016) to generate high-quality chest X-ray diagnostic reports that closely resemble those written by radiologists. Wu et al. conducted a study evaluating the performance of the GPT-4V(ision) (Yang et al., 2023b) model in the context of multimodal medical diagnostics (Wu et al., 2023a). The assessment covered 17 different organ systems, including the central nervous system, head and neck, cardiac, thoracic, hematologic, hepatobiliary, gastrointestinal, urogenital, gynecological, obstetrical, breast, musculoskeletal, spinal, and vascular systems. The images utilized were derived from common clinical modalities such as radiographs, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), digital subtraction angiography (DSA), mammograms, ultrasound scans, and pathology slides. While the study highlighted the model's strong ability to identify imaging modalities and anatomical structures, it also revealed significant limitations in its capacity to provide accurate disease diagnoses and generate comprehensive medical reports. Li et al. proposed a model called ASGK (Auxiliary Signal-Guided Knowledge Encoder-Decoder), which utilizes Generative Pre-Training (GPT) (Radford, 2018) as its natural language decoder to generate medical reports (Li et al., 2023b). Li et al. used the MIMIC-CXR dataset (Johnson et al., 2019a,b) and adopted zero-shot and few-shot prompting strategies to evaluate the performance of GPT-4V in chest X-ray report generation tasks (Li et al., 2023c). The experimental results demonstrated that GPT-4V showed significant potential in medical report generation, with the mixed prompt (Few-shot mixed examples prompt) producing the highest-quality reports, highlighting the importance of combining different types of prompts to improve performance. However, further optimization of prompt design and data training remains critical, especially as GPT-4V still exhibits limitations in specific disease descriptions and certain evaluation metrics, such as CIDEr. Li et al. proposed a model named KARGEN (Knowledge-Enhanced Automated Radiology Report Generation), which leverages a framework combining LLaMA with a medical domain knowledge graph (Li et al., 2024b). The model utilizes a pre-trained Swin Transformer to extract regional image features, representing local information in chest X-ray images and the LLaMA2-7B generates diagnostic reports using the fused visual features as prompts. In other experiments (Bannur et al., 2023a; Boecking et al., 2022; Delbrouck et al., 2022), pretrained language models have also been utilized in research on ARRG.

In summary, foundation models have significantly advanced the field of ARRG. Their ability to integrate visual and textual data, coupled with their scalability and adaptability, has enabled the generation of clinically relevant and structured diagnostic reports. However, challenges remain, particularly in improving the accuracy of disease-specific descriptions and optimizing prompt design for these foundation models. Meta Prompting, proposed by Zhang et al., is a structured and syntax-oriented prompting method designed to enhance the reasoning capabilities of foundation models without relying on specific examples (Zhang et al., 2023b). Unlike few-shot prompting, which primarily depends on learning from examples, Meta Prompting emphasizes the general structure of tasks, breaking down complex tasks into structured steps to improve the consistency and stability of the generation process. Based on this, Meta Prompting holds potential applications in ARRG research, as it can be leveraged to optimize the structural coherence of generated reports, enhancing their consistency and readability. Moreover, it can improve token efficiency, reduce redundant computations during report generation, lower computational costs, and ultimately enhance the clinical utility of radiology reports.

5.1.5. Hybrid models

Hybrid models are treated here as a generation paradigm because they combine multiple report-generation logics within a single overall pipeline, such as retrieval guidance or template guidance plus neural decoding. Unlike the strategies discussed later, the emphasis in this subsection is on hybrid systems whose full report-generation workflow is itself structurally mixed.

Li et al. proposed a model called Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) for medical imaging report generation (Li et al., 2018). The HRGR-Agent model integrates retrieval-based and generation-based methods and is optimized using reinforcement learning (RL). The model follows a hierarchical decision-making process, where the retrieval policy module determines whether to select a sentence from a template database. If no suitable template is found, the Generation Module is invoked to generate a new sentence. The HRGR-Agent is trained through reinforcement learning and utilizes sentence-level and word-level reward mechanisms to achieve ARRG. Later, Li et al. proposed a knowledge-driven framework named KERP (Knowledge-driven Encode, Retrieve, Paraphrase) (Li et al., 2019a). The KERP model consists of three key modules: encode, retrieve, and paraphrase. Specifically, the model first employs DenseNet (Huang et al., 2017) to extract visual features from medical images. Then, Graph Transformer (GTR) further processes these features by converting them into a structured abnormality graph, which represents potential medical abnormalities and is optimized using prior medical knowledge. In the text generation phase, KERP adopts both retrieve-based and template-based approaches. The retrieval phase, using GTRg2s (Graph Transformer from Graph to Sequence), the model retrieves the most relevant template sequence from a predefined medical report template repository based on the abnormality graph. The paraphrasing phase, utilizing GTRgs2s (Graph Transformer from Graph & Sequence to Sequence), the retrieved templates are refined by incorporating information from the abnormality graph, ultimately generating the final medical report. Wang et al. proposed a unified framework called Relation-paraNet for retrieval and

relational topic-driven sentence generation (Wang et al., 2020). This framework combines template retrieval and sentence generation to dynamically decide whether to retrieve common descriptions from a template library or generate new sentences, thereby effectively handling both common and rare abnormalities. It incorporates a Relational Abnormality Classification Module to identify medical terms and their semantic relationships, ensuring that the classification results of abnormal terms are consistent with their actual relationships. Meanwhile, the relational-topic encoder integrates image features with contextual information to generate topic vectors that ensure the semantic coherence of sentence generation. Finally, the Adaptive Generator dynamically switches between template retrieval and sentence generation modes based on the context, guiding the creation of medical reports. In Liu et al.'s study, they proposed the PPKED model (Liu et al., 2021a). In the text generation phase, posterior knowledge is leveraged to retrieve prior knowledge. The Prior Knowledge Explorer (PrKE) utilizes posterior knowledge as a query to retrieve relevant textual information from the historical report database and the medical knowledge graph. Additionally, the study introduces the Adaptive Distilling Attention (ADA) mechanism, which dynamically adjusts the weighting of posterior and prior knowledge during report generation. Finally, all this knowledge is integrated using the Transformer decoder, resulting in a comprehensive radiology report. Yang et al. incorporated retrieval-augmented specific medical knowledge to accomplish the text generation task (Yang et al., 2022). First, they computed the disease label distribution of the given X-ray image and retrieved the most similar historical radiology reports from the database based on KL-Divergence. Then, they employed Stanza (Qi et al., 2020; Zhang et al., 2021) to extract medical terms from the retrieved reports and queried the RadGraph (Jain et al., 2021) knowledge graph to construct knowledge triplets relevant to the current case. These knowledge representations were semantically encoded using Clinical-BERT (Alsentzer et al., 2019), converting the medical terms obtained from RadGraph (Jain et al., 2021) into knowledge vectors. These vectors were subsequently integrated with visual features and general knowledge through the knowledge-enhanced multi-head attention (KEMHA) to provide medical contextual information for text generation. Finally, these fused knowledge vectors, along with visual features and general knowledge, were processed through the KEMHA before decoding to ensure that the text generation process effectively leveraged medical knowledge. The integrated information was then fed into a standard decoder to generate the final radiology report. Some other studies (Han et al., 2018, 2021; Xie et al., 2019) have also employed hybrid methods to explore the generation of ARRG.

Table 7 summarizes the main methodological paradigms across a shared set of analytical dimensions, including input organization, prior-knowledge use, support for long-range or paragraph-level generation, control over report structure and factual consistency, and typical failure modes. The citations embedded within individual cells provide the supporting studies and review evidence for each comparative characterization.

5.2. Enhancement mechanisms

5.2.1. Attention mechanisms

With the advancement of deep learning, many studies have adopted the CNN-RNN framework in the field of image captioning (Vinyals et al., 2015). Since Bahdanau et al. introduced the attention mechanism (Bahdanau, 2014), an increasing number of studies have incorporated visual attention (Donahue et al., 2015; Xu, 2015) into the CNN-RNN framework, while some have also explored semantic attention (You et al., 2016). In ARRG, attention should not be understood only as an optional add-on. In early CNN-RNN systems it often appeared as a module, whereas in Transformer-based and foundation-model-based systems it became a core representational mechanism.

A similar evolution can be observed in ARRG, which, while sharing similarities with image captioning, presents unique challenges due to the critical nature of medical diagnosis and the need for precise and informative descriptions. Early ARRG systems often treated the entire image as a global feature vector, neglecting the rich spatial information present in radiological images. These systems, including early CNN-RNN encoder-decoder pipelines and compositional approaches, often overlooked spatial detail that is crucial for generating accurate descriptions. In contrast, attention-based techniques dynamically identify key regions of the input image during output sequence generation, enabling models to capture finer-grained features and contextually relevant details (Reale-Nosei et al., 2024). This has led to the widespread adoption of attention mechanisms in ARRG, either as modules in encoder-decoder systems or as core components in Transformer-based models. Jing et al. were the first to introduce the co-attention into the field of ARRG (Jing et al., 2017). They proposed a co-attention model that integrates visual and semantic features, achieving dynamic alignment between medical image regions and semantic labels. This mechanism employs visual attention to locate critical image regions (such as abnormalities) and semantic attention to capture high-level semantic information (such as diagnostic labels), ultimately generating a joint context vector for paragraph generation. Wang et al. proposed the TieNet model, which employs a multi-level attention mechanism (including visual attention and text-based attention) for disease classification and report generation from chest X-rays (Wang et al., 2018). Additionally, it is a complete end-to-end model (CNN-RNN) that integrates the attention mechanism with multi-task learning. Xue et al. proposed a multimodal recurrent model. In this model, the attention mechanism is primarily applied during the text generation phase, helping the model dynamically focus on different regions of the medical image while generating each sentence (Xue et al., 2018). Specifically, whenever a new sentence is generated, it relies not only on the semantic information of the previous sentence but also on visual attention to select the most relevant regions of the image. For example, when describing the heart, the attention mechanism enhances focus on the heart region, whereas when describing the lungs, it readjusts its focus accordingly. Yin et al. incorporated the attention mechanism into the sentence-level RNN (Yin et al., 2019). When the sentence-level RNN generates topic vectors, it calculates attention weights for each image region and tag. These weights indicate which regions of the image and tags the model should focus on when generating the current sentence. By leveraging the attention mechanism, the model can better capture local information in the image that is relevant to the current sentence. You et al. proposed the Align Hierarchical Attention mechanism and integrated it into the AlignTransformer model (You et al., 2021). This mechanism leverages a hierarchical attention approach to precisely align visual features from medical images with disease labels, thereby enhancing the accuracy and semantic consistency of diagnostic report generation. By incorporating this alignment mechanism, the AlignTransformer model achieves finer-grained mapping in the dynamic association between visual and semantic features. Song et al. proposed the CMCA model (Cross-modal Contrastive Attention Model), which leverages a historical database of similar cases to extract unique abnormal regions from input images through a contrastive mechanism (Song et al., 2022). Simultaneously, it retrieves semantic information related to these abnormalities from the corresponding reports of similar cases. By dynamically aligning visual features with semantic features, the model effectively guides the generation of medical imaging reports. Gajbhiye et al. proposed the AMLMA model (Adaptive Multilevel Multi-Attention), an adaptive multilevel multi-attention mechanism designed for medical image report generation (Gajbhiye et al., 2022). The model introduces a multilevel attention mechanism to dynamically align visual features with semantic features, integrating Adaptive Visual-Semantic Attention (AVSA) and Visual-Based Language Attention (VBLA). These mechanisms are used respectively to extract key regions in medical images and dynamically adjust semantic weights during report generation. Additionally, the model incorporates a Residual Attention Module (RAM) to enhance multi-label abnormality detection capabilities. Wang et al. proposed an innovative model called Memory-augmented Sparse

Table 7
Cross-study synthesis of major ARRG methodological paradigms across shared analytical dimensions.

Paradigm	Typical Input / View Organization	Prior Knowledge Incorporated?	Long-Range Context / Paragraph-Level Generation	Granularity and Control over Report Ordering / Consistency	Typical Failure Modes
Template-Based Models	Usually single-modality input; fixed slots or predefined section layout (Abela et al., 2022), (Pino et al., 2021)	Yes, but mainly as handcrafted templates or reporting rules (Abela et al., 2022), (Pino et al., 2021)	Limited; usually sentence- or slot-level rather than flexible paragraph generation (Abela et al., 2022), (Messina et al., 2022)	High control over wording, ordering, and required report elements (Abela et al., 2022), (Pino et al., 2021)	Rigid outputs; poor adaptation to rare or complex cases; template mismatch (Beddiar et al., 2023), (Messina et al., 2022)
Retrieval-Based Models	Typically single-image or multi-view similarity matching against a report bank (Zhang et al., 2018), (Yang et al., 2021c)	Implicitly yes, through stored reports, retrieved cases, or sentence libraries (Ni et al., 2020), (Yang et al., 2021c)	Moderate; can reuse paragraph-level text, but does not truly model long-range generation (Kougia et al., 2021), (Yang et al., 2021c)	Moderate control if retrieval is good; ordering often inherited from retrieved exemplars (Zhang et al., 2018), (Yang et al., 2021c)	Database dependence; limited novelty; semantically plausible but case-mismatched reports (Beddiar et al., 2023), (Liu et al., 2025)
Encoder-Decoder Paradigms	Usually image-to-text; supports single-view, multi-view, and some multimodal extensions (Shin et al., 2016b), (Yuan et al., 2019), (Liu et al., 2021a)	Optional; may incorporate labels, tags, graphs, or retrieved context, but not always (Jing et al., 2017), (Zhang et al., 2020), (Liu et al., 2021a)	Yes; especially with hierarchical decoders, attention, or Transformer components (Jing et al., 2017), (Yin et al., 2019), (You et al., 2021)	Moderate control; stronger than free generation when guided, but still prone to ordering and section-consistency issues (Yuan et al., 2019), (You et al., 2021)	Long-range dependency problems, repetition, omission of abnormalities, section inconsistency, factual drift (Liao et al., 2023), (Kaur et al., 2022), (Sloan et al., 2024)
Foundation-Model-Based Approaches	Can support broader multimodal and multi-view inputs; often built on pretrained vision-language backbones (Nooralahzadeh et al., 2021), (Zhou et al., 2022), (Wu et al., 2023a)	Often yes, through large-scale pretraining, prompts, external knowledge, or aligned corpora (Alfarghaly et al., 2021), (Li et al., 2024b), (Nicolson et al., 2023)	Strong in principle for long-range context and paragraph-level generation (Chen et al., 2020), (Alfarghaly et al., 2021), (Nooralahzadeh et al., 2021)	Tends to offer lower intrinsic control unless constrained by prompting, retrieval, templates, or additional supervision (Li et al., 2023c), (Wu et al., 2023a), (Li et al., 2024b)	Hallucination, overconfident fluent errors, weak rare-abnormality grounding, high data/computation demands (Li et al., 2023c), (Wu et al., 2023a), (Sloan et al., 2024)
Hybrid Models	Can combine image input, retrieved reports, templates, prior studies, and multimodal context in one pipeline (Wang et al., 2020), (Liu et al., 2021a), (Yang et al., 2022)	Frequently yes; often combines retrieval, templates, or structured medical knowledge (Li et al., 2019a), (Liu et al., 2021a), (Yang et al., 2022)	Potentially strong, depending on how retrieval and generation are coordinated (Li et al., 2018), (Wang et al., 2020)	Often offers better controllability than pure generation while preserving more flexibility than templates alone (Wang et al., 2020), (Yang et al., 2022), (Nie & Liu, 2023)	Pipeline complexity; error propagation between modules; higher engineering and data requirements (Li et al., 2018), (Liu et al., 2021a), (Nie & Liu, 2023)

Attention (MSA), which combines bilinear pooling with self-attention mechanisms to capture high-order interactions among features (Wang et al., 2022b). The model introduces memory slots to encode and store accumulated features throughout the processing stages, enhancing contextual understanding and generation capabilities for long-form reports. In generating medical imaging reports, the MSA module uses global features as queries and local features as keys and values, leveraging sparse attention to efficiently extract relevant information, thereby producing more accurate and coherent long-text reports. Yan et al. proposed a model based on a two-level sparse attention mechanism, aimed at enhancing semantic alignment in medical image report generation (Yan et al., 2022). The model employs Region Sparse Attention to extract key regional features from medical images corresponding to each MeSH (FB, 1963) label, while Word Sparse Attention assigns higher attention weights to MeSH (FB, 1963) terms present in the report, ensuring dynamic alignment between visual and semantic features.

5.2.2. Multimodal fusion

Radiologists, when writing imaging diagnostic reports, do not simply "describe what they see" but first review the patient's medical history, previous examination records, and laboratory results to provide essential contextual support for image interpretation (Wang et al., 2024b). For instance, in patients with tumours, special attention is paid to the presence of metastatic lesions. Similarly, when a pulmonary mass is identified on CT but presents atypically, a positive mycobacterial culture strongly suggests tuberculosis rather than lung cancer. Inspired by this practical workflow, researchers have proposed integrating multimodal methods into ARRG (Dalla Serra et al., 2022; Jeong et al., 2024; Mondal et al., 2023; Nguyen et al., 2023, 2021; Shang et al., 2022).

In Nguyen et al.'s study, chest X-ray images and patients' clinical history documents are combined through a classification module to generate disease embeddings (Nguyen et al., 2021). These embeddings are then passed to a Transformer-based generator to produce medical reports. Simultaneously, the generator outputs a weighted embedding representation, which is fed into an interpreter to ensure consistency with disease-related topics. In the experiments conducted by Dalla Serra et al., clinically relevant structured information is extracted from images in a supervised manner, represented in the form of triples, and these triples are used as input for generating radiology reports (Dalla Serra et al., 2022). In the experiments conducted by Shang et al., a model named Multimodal Adaptive Transformer (MATNet) was proposed, which integrates radiological images and clinical records to enhance the learning capabilities of its encoder (Shang et al., 2022). Mondal et al. proposed a model named TransXpainNet, which utilizes a domain-knowledge-based vision transformer, DeiT-CXR, to extract image features (Mondal et al., 2023). Additionally, clinical history and other documents are incorporated to enrich the report generation process. Although much of the early multimodal literature focused on 2D chest X-rays, the same motivation extends naturally to 3D imaging. In CT and MRI, multimodality is not only about adding text-side clinical context; it is also about integrating volumetric image context, prior examinations, phase information, and sometimes multiple reconstruction windows. This introduces technical challenges such as the curse of dimensionality, slice aggregation or fusion strategy design, and the high computational cost of 3D encoders, which partly explains why multimodal ARRG for 3D data remains less mature than chest X-ray report generation.

5.2.3. Knowledge integration

Knowledge integration is discussed here as an enhancement mechanism rather than as a separate core paradigm. The focus is therefore not on whether a system is hybrid, retrieval-based, or encoder-decoder in overall form, but on how external medical knowledge, knowledge graphs, prior reports, or clinically structured signals are injected to improve factual grounding and report quality.

Babar et al. examined the effectiveness of encoder-decoder models and noted that current ARRG models predominantly operate on unconditional generation patterns, which do not adequately leverage image features for producing high-quality reports (Babar et al., 2021a). They recommended integrating medical knowledge into model design to enhance the quality and precision of generated reports (Babar et al., 2021a). Some researchers have proposed methods to incorporate medical knowledge into models (Gajbhiye et al., 2022; Hou et al., 2023b; Kim et al., 2023; Li et al., 2022; Nishino et al., 2022), while others have developed models based on knowledge graphs (Li et al., 2024b; Wang et al., 2023a; Zhang et al., 2020). Kim et al. argued that models like Transformers often lack prior knowledge, leading to errors such as referencing non-existent previous examinations in the generated reports (Kim et al., 2023). This discrepancy can be attributed to the knowledge gap between radiologists and the generative models. To address this issue, they introduced a rule-based tagger to extract comparative prior information from radiology reports. This extracted prior information was then integrated into two publicly available models (R2Gen Chen et al., 2020 and M^2Tr Cornia et al., 2020), enabling them to generate more accurate and comprehensive reports. Additionally, similar reports are retrieved from a pre-built repository based on the visual appearance of the input image, further enhancing the quality of report generation. Zhang et al. proposed a model that integrates knowledge graphs and deep learning techniques for radiology report generation (Zhang et al., 2020). The input to the decoder is a context vector extracted through a graph attention mechanism from the graph embedding module. Yang et al. proposed a chest radiology report generation framework that integrates both general and specific knowledge (Yang et al., 2022). The framework extracts broad medical knowledge from a knowledge graph while retrieving specific knowledge related to the current input image. It employs a knowledge-enhanced multi-head attention mechanism to combine these knowledge components with visual features from the image, enabling the generation of detailed radiology reports. Subsequently, in their later work (Yang et al., 2023a), they introduced an automated radiology report generation model that integrates knowledge-driven and multimodal alignment mechanisms. The model features a dynamically updated knowledge base, which autonomously learns and stores medical knowledge throughout the training process. Specifically, the model extracts visual features from images and textual embeddings from corresponding reference reports during training, iteratively updating the knowledge base. During inference, this fixed knowledge base is combined with image features to produce clinically relevant radiology reports.

5.2.4. Reinforcement learning

Reinforcement learning is likewise treated here as a cross-cutting optimization strategy rather than as a separate report-generation paradigm. Its role is to refine generation behaviour through reward design, and it may be combined with hybrid, encoder-decoder, or other model families. For this reason, some individual studies appear both conceptually hybrid and reinforcement-learning-based, but the present subsection focuses specifically on the optimization role of reinforcement learning.

For example, in the experiment conducted by Li et al., HRGR-Agent employs Hierarchical Reinforcement Learning (HRL) for policy optimization, incorporating both sentence-level and word-level reward mechanisms (Li et al., 2018). The model uses CIDEr as the reward function and applies the reinforce algorithm to optimize the retrieval policy module and the generation module separately. The reward for the retrieval policy module is based on the sentence-level CIDEr incremental score, which determines whether to retrieve a template or generate a new sentence. Meanwhile, the generation module is optimized using the word-level CIDEr incremental score, refining the generation of each word. Liu et al. employed reinforcement learning for report generation (Liu et al., 2019). Their approach incorporated natural language generation (NLG) metrics and clinical accuracy as reward signals. In the study by Jing et al., reinforcement learning was used to optimize text generation within the CMAS model (Jing et al., 2020). The model was first

initialized using supervised learning. Subsequently, the reinforce algorithm was applied to directly optimize BLEU-4 and CIDEr evaluation metrics. Through a discounted reward mechanism, the model enhanced the focus of the Abnormality Writer (AW) on abnormalities, thereby improving the clinical utility of the generated radiology reports. Additionally, other studies, such as Xiong et al. (2019), have explored the application of reinforcement learning in ARRG.

Collectively, these studies indicate that reinforcement learning can enhance the fluency and clinical accuracy of automatically generated radiology reports, thereby improving their clinical utility.

Taken together, the differences among the methods reviewed in this section lie not only in their architectures, but also in their input organization, use of prior knowledge, support for long-range or paragraph-level generation, degree of control over report ordering and factual consistency, and vulnerability to different failure modes. Template-based and retrieval-based methods are generally more controllable and often perform relatively well in standardized or repetitive reporting scenarios, but their flexibility is often limited when reports need to describe complex or rare findings. In contrast, encoder-decoder methods, Transformer-based methods, and more recent foundation-model-based approaches are better suited to richer visual inputs and paragraph-level report generation; however, their greater flexibility may also introduce a higher risk of failure modes such as hallucination, omission of clinically important findings, and section inconsistency. Hybrid methods attempt to balance these trade-offs by combining stronger controllability with greater expressive flexibility. At the same time, better surface-level natural language generation performance does not necessarily imply greater clinical usefulness, especially when the generated reports remain highly repetitive or overly biased toward generic normal-case descriptions. These differences are therefore important for understanding and interpreting comparative performance results, and they are revisited in the next section.

6. Evaluation methods

The goal of ARRG is to produce reports that align closely with those written by radiologists. Therefore, effectively evaluating ARRG models is crucial. However, there is currently no unified evaluation metric for ARRG models. Since most experimental datasets include either corresponding reports, annotations, sentence-level descriptions, labels, and other related information, existing studies primarily employ both quantitative evaluation metrics and qualitative evaluation methods.

6.1. Quantitative evaluation methods

6.1.1. Natural language generation (NLG) evaluation metrics

NLG evaluation metrics primarily assess the linguistic quality of generated reports, such as fluency, text coverage, and structural consistency. These methods evaluate the similarity between the generated and reference descriptions based on word overlaps, measuring how many words or n-grams (phrases with n consecutive words) are shared between them.

- *Bilingual Evaluation Understudy (BLEU):* BLEU (Papineni et al., 2002) is based on n-gram precision, used to evaluate the lexical and word-order match between generated and reference texts. It quantifies the precision of the generated text by calculating n-gram overlaps. BLEU is typically presented in the form BLEU-n, where n represents the n-gram length (e.g., BLEU-1 for unigram matches, BLEU-4 for matches up to four-grams). A brevity penalty is introduced to penalize overly short descriptions. The closer the BLEU score is to 1, the better the model performance. The BLEU metric is easy to compute and interpret, and it has shown strong correlation with human judgments when assessing the quality of generated text. However, BLEU focuses solely on lexical matching between the generated and reference texts, failing to capture semantic consistency or the logical coherence of the text (Pang et al., 2023).
- *Metric for Evaluation of Translation with Explicit ORdering (METEOR):* METEOR (Banerjee & Lavie, 2005) is metric that calculates precision and recall for unigram matches between generated and reference texts, combining them into an F-score. METEOR shares BLEU's limitation of not considering the coherence or overall quality of generated text (Pang et al., 2023).
- *Recall-Oriented Understudy for Gisting Evaluation (ROUGE):* ROUGE (Lin, 2004) is a set of recall-based metrics initially designed for text summarization but widely used in natural language generation tasks. It evaluates the coverage of generated content by measuring overlaps in n-grams, longest common subsequences (LCS), and other aspects between generated and reference texts. Common versions include ROUGE-N, ROUGE-L, and ROUGE-W, with ROUGE-L focusing on longest common subsequence matching. However, ROUGE-L only considers a single aspect-longest common subsequence-which may fail to capture all dimensions of text quality (Pang et al., 2023).
- *Consensus-based Image Description Evaluation (CIDEr):* CIDEr (Vedantam et al., 2015) is an evaluation metric specifically designed for image description tasks. It calculates the cosine similarity between the generated and reference texts in terms of n-gram overlaps weighted by TF-IDF scores. This weighting emphasizes the importance of content-specific n-grams in the evaluation process.
- *BERTScore:* BERTScore (Zhang et al., 2019a) is an automated evaluation method for text generation. This metric leverages contextual embeddings to compute the semantic similarity scores between individual words in the generated and reference sentences. The study by Yu et al. indicates that automated metrics, such as BERTScore, demonstrate better alignment with radiologists' assessments (Yu et al., 2022). In recent years, the ImageCLEFmedical challenge has adopted BERTScore as one of its evaluation metrics (Ionescu et al., 2023).

6.1.2. Clinical efficacy (CE) evaluation metrics

NLG evaluation metrics are primarily used to measure the similarity between generated reports and reference reports, but they do not necessarily reflect the medical facts contained within the reports accurately (Babar et al., 2021b; Boag et al., 2020; Liu et al., 2019; Pino et al., 2021, 2020; Zhang et al., 2019b). For example, the sentences 'Effusion' observed and 'No effusion' observed may appear very similar on the surface. Metrics based on n-gram matching might assign a high score to such pairs, even though their meanings are entirely opposite in terms of medical facts. Another issue is the growing perspective that it is more important to assign higher scores to reports that are semantically equivalent (Babar et al., 2021a; Boag et al., 2020). For instance, the sentences 'The heart is within normal size and contour and No observed cardiomegaly.' express the same meaning but, due to differences in wording, might receive a score of 0 under commonly used NLG metrics. To address this issue, the combination of CE methods and NLG techniques is being increasingly and widely adopted.

- *Precision, Recall and F1-score:* The CheXpert dataset (Irvin et al., 2019) labeling tool was used to extract and analyse labels from both the generated reports and the ground truth reports. By comparing the labels from the generated reports with those from the ground truth reports, the study employed metrics such as precision recall and F1-score to evaluate the performance of the ARRG model (Liu et al., 2019). These evaluation metrics have been widely applied in ARRG research (Chen et al., 2020; Liu et al., 2019; Lovelace & Mortazavi, 2020; Nooralahzadeh et al., 2021).
- *MeSH Accuracy (MA):* The ratio of the number of correctly generated MeSH (FB, 1963) terms in a model-generated report to the total number of MeSH (FB, 1963) terms present in the reference report.
- *Keyword Accuracy (KA):* The ratio of the number of correctly generated keywords in the generated report to the total number of keywords in the reference report. The keyword set is constructed from MTI (Medical Text Indexer) annotations of the original dataset and

manual annotations, containing 438 unique medical keywords (Xue et al., 2018).

- *Medical Image Report Quality Index (MIRQI):* It evaluates the accuracy of disease keywords and their associated attributes (e.g., lesion location, severity, negation, or uncertainty) in generated reports compared to ground truth reports through a graph-based matching approach. MIRQI defines three core metrics: Recall (MIRQI-r), Precision (MIRQI-p), and F1 Score (MIRQI-F1), which collectively assess the performance of report generation in terms of disease mentions and attribute alignment (Zhang et al., 2020).
- *Radiology Report Quality Index (RadRQI):* In the experiments conducted by Yan et al., a novel evaluation method called Radiology Report Quality Index (RadRQI) was proposed (Yan et al., 2023). This method assesses the alignment between generated reports and ground truth reports by extracting abnormalities and their associated attributes (such as anatomical location or condition) and establishing relationships between them. RadRQI leverages RadLex (Langlotz, 2006) to extract relevant keywords and utilizes RadGraph (Jain et al., 2021) to contextualize the associations between abnormalities and attributes, while also accounting for negations to accurately capture the presence or absence of abnormalities mentioned in the reports. The core metrics of RadRQI include RadRQI-F1, which measures the correctness of the abnormalities and their attributes in the generated reports, and RadRQI-Hits, which evaluates the coverage of distinct abnormality categories in the generated reports.
- *RadGraph F1 & RadCliQ:* Yu et al. proposed two evaluation metrics for radiology report generation: RadGraph F1 and RadCliQ (Yu et al., 2022). RadGraph F1 evaluates the quality of generated reports by comparing the overlap of clinical entities (e.g., pathologies and anatomical locations) and their relationships extracted from machine-generated and reference reports. This metric introduces the concept of knowledge graphs, structuring diagnostic information into entities and their relationships to more accurately capture the semantic and clinical relevance of the reports. RadCliQ, on the other hand, is a composite evaluation metric that integrates multiple individual metrics (such as BLEU, BERTScore, CheXbert, and RadGraph F1) and assigns weights to each through a linear regression model, providing a more accurate reflection of radiologists' assessments of the clinical quality of the reports. These two metrics aim to address the limitations of traditional evaluation methods, offering more clinically meaningful standards for assessing generated reports.

6.2. Qualitative evaluation methods

Qualitative evaluation methods primarily rely on human experts, such as radiologists, to subjectively assess the quality of generated reports.

- *Human evaluation:* The researchers conducted a quality evaluation of the reports generated by each method through the Amazon Mechanical Turk platform. They randomly selected 100 test samples and provided ground truth reports as references. Participants were invited to choose the report generated by different models that best matched the reference report. The evaluation criteria included language fluency, content selection relevance, and the correctness of medical abnormal findings. A default option was provided to handle cases where participants had no preference or preferred both reports (Li et al., 2019a, 2018). In the experiment conducted by Alfarghaly et al., human evaluation was also performed by a radiologist with five years of experience in interpreting chest X-ray images (Alfarghaly et al., 2021). By comparing the automatically generated reports with the ground-truth reports, the radiologist categorized the generated reports into three types: accurate reports, reports with missing details, and false reports. Grad-CAM visualization technology (Selvaraju et al., 2017) was used to assist in the evaluation process. In another experiment, Gale et al. mentioned two methods of human evaluation: sentence Content and acceptance of explanations by doctors. The former involves a radiologist reviewing the generated sentences and original report sentences for 200 randomly selected fracture cases, assessing whether they accurately describe the location and characteristics of the fractures. The latter involves inviting five doctors with 3 to 7 years of postgraduate clinical experience to evaluate 30 fracture cases from the test set, scoring three different forms of explanations: saliency maps, generated sentences and a combination of saliency maps and generated sentences (Gale et al., 2019).
- *Likert Scale Scores:* A scoring system designed to evaluate medical experts' opinions on the quality of generated reports (Fleiss, 1971).

Taken together, quantitative and qualitative evaluation methods capture different but complementary aspects of ARRG quality. Quantitative metrics support scalable comparison across models, but they do not fully capture factual correctness, logical consistency, or clinical usefulness. Qualitative evaluation remains more clinically informative, yet it is time-consuming, costly, and difficult to scale. Table 8 summarizes these trade-offs. Overall, future progress will depend on more balanced evaluation frameworks that combine efficient automatic metrics with clinically grounded human or semi-automated assessment.

6.3. Comparison of ARRG performance

In this part, we compare the experimental results of ARRG studies, including NLG evaluation metrics and CE evaluation metrics, to identify the most advanced methods. Currently, IU X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019a,b) are the most commonly used datasets in ARRG research. To ensure a fair comparison, our analysis is based on these two datasets. All cited numerical values are derived from the original papers to ensure accuracy and comparability. Tables 9 and 10 present a comparative evaluation of different research methods based on the IU X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019a,b) datasets, respectively.

In the experimental results presented in Table 9, RadioBERT (Kaur, 2022) demonstrated outstanding performance on the IU X-ray dataset (Demner-Fushman et al., 2016). In this experiment, the integration of DistilBERT not only provided context-aware word embeddings, enhancing the semantic coherence of the generated text, but was also utilized for sentiment analysis, enabling sentence reordering so that abnormal descriptions were prioritized. As a lightweight version of BERT, DistilBERT reduces model parameters and computational overhead while maintaining strong semantic understanding, thereby improving text generation efficiency. Additionally, the experiment employed HLSTM for text generation, which effectively captures the hierarchical structure of radiology reports, enhancing the logical consistency of the text. This suggests that foundation models can be combined with other generation models (such as HLSTM) to optimize computational efficiency while further improving the overall performance of ARRG models.

Furthermore, Table 9 also shows that the study conducted by Kale et al. (2023) achieved excellent results in NLG evaluation metrics on the IU X-ray dataset (Demner-Fushman et al., 2016). This may be attributed to the fact that NLG evaluation metrics primarily stem from their reliance on n-gram matching and word order similarity to assess the resemblance between the generated and reference texts. Since template-based methods employ predefined sentence structures and fixed expressions, the generated reports exhibit a high degree of similarity to reference reports in terms of vocabulary selection, phrase order, and syntactic structure, thereby achieving higher scores in NLG evaluation metrics.

From Table 10, it can be observed that the study by Kale et al. (2023) on ARRG also attained the best performance in NLG evaluation metrics on the MIMIC-CXR dataset (Johnson et al., 2019a,b). This may be due to the use of predefined sentence structures and fixed expressions in template-based methods, ensuring that the generated reports closely match the reference reports in terms of word choice, phrase order, and

Table 8
Comparison of ARRG evaluation methods: advantages and disadvantages.

Method	Advantages	Disadvantages
BLEU	Easy to implement	Fail to capture semantic consistency
METEOR	Combines precision and recall; Considers synonyms and stemming	Fail to capture deep semantics
ROUGE	Suitable for long texts; multiple variants (e.g., ROUGE-L) assess different aspects	Fail to reflect semantic consistency
CIDEr	Uses TF-IDF weighting to emphasize important words	Fail to capture medical fact accuracy
BERTScore	Captures semantic similarity	High computational cost; Relies on pre-trained models
Precision, Recall, F1	Directly evaluates medical label accuracy	Requires high-quality annotated data
MA	Evaluates accuracy of medical terms	Depends on MeSH term coverage
KA	Evaluates keyword accuracy	Relies on keyword set quality; Fail to assess semantic depth
MIRQI	Evaluates disease keywords and attributes	High computational complexity; Requires quality annotated data
RadRQI	Evaluates abnormal findings and attributes; Uses RadLex and RadGraph for comprehensive assessment	High computational complexity; Requires medical knowledge
RadGraph F1	Combines knowledge graphs; Integrates RadRQI for	High computational complexity; Requires quality annotations
& RadCliQ	comprehensive evaluation	and medical knowledge
Human Evaluation	Most reliable; Assesses language quality, medical factual accuracy, and clinical usability	High cost; Time-consuming; Difficult to scale; Subjective bias
Likert Scale Scores	Quantifies expert opinions; provides subjective evaluation	Depends on evaluator expertise; Potential variability between evaluators

syntactic structure. On the other hand, compared to natural datasets, the MIMIC-CXR dataset (Johnson et al., 2019a,b) exhibits a certain degree of bias. Specifically, the dataset contains fewer abnormal cases than normal ones, and the complexity of the chest X-ray cases is relatively low. As mentioned earlier, template-based methods tend to perform well when handling normal images or common diseases. Under this data distribution, evaluating model performance using NLG metrics may lead to an unfair conclusion. As a result, template-based methods tend to excel in NLG evaluation metrics. However, it is important to note that although template-based methods achieve high scores in NLG evaluation, their adaptability in real-world clinical applications is relatively limited, particularly when handling complex or rare cases, where they may face challenges in expressive flexibility.

In terms of CE evaluation metrics, the experimental results in Table 9 indicate that texts generated using LSTM or Transformer models often outperform traditional template-based and retrieval-based methods. This is likely because template-based methods are constrained by predefined sentence structures, making it difficult to accurately describe cases that fall outside the predefined templates. Also, retrieval-based methods depend on existing case repositories for matching, which may result in reports that do not fully correspond to the specific details of rare or complex cases. In contrast, LSTM and Transformer models offer greater flexibility in adapting to different case scenarios, generating more personalized and clinically relevant reports, thereby achieving superior performance in CE evaluation metrics.

7. Ethical issues

Research on ARRG is attracting increasing attention from scholars and researchers. Its core objective is to alleviate the workload of radiologists, improve the accuracy and efficiency of report generation, and ultimately provide better medical services for patients. However, the application of ARRG also raises a series of ethical concerns, primarily involving three aspects: patients, doctors, and regulatory authorities (Char et al., 2018; Kaissis et al., 2020). Beyond the general concerns discussed in medical AI governance, ARRG also introduces task-specific risks because it generates long-form clinical text. A report may appear fluent and plausible while still containing unsafe content, such as hallucinated findings, omission of key abnormalities, errors in handling negation, incorrect references to prior examinations, or inconsistencies between the Findings and Impression sections. In addition, models may overproduce templated normal descriptions, thereby masking rare abnormalities or making clinically important findings appear less salient. These failure modes are particularly important because they are difficult to detect from language fluency alone and can directly affect downstream clinical decision-making. From an ethical perspective, such failures are not merely technical imperfections; they bear directly on patient safety, informed clinical judgment, responsibility attribution, and the trustworthiness of AI-assisted reporting in real clinical workflows.

7.1. Patients: privacy, security, and autonomy

From the patient's perspective, the widespread application of ARRG mainly involves issues related to privacy and data security, algorithmic bias and fairness, transparency and explainability, as well as informed consent and patient autonomy.

7.1.1. Privacy and data security

ARRG systems require vast amounts of medical imaging data for training, which often contain patients' Protected Health Information (PHI), such as names, ages, contact details, addresses, and medical histories. If such data is compromised due to hacking or poor management, it could trigger public concern and diminish trust in AI-assisted healthcare. Ensuring patient privacy and data security is, therefore, of utmost importance. To address this issue, federated learning can be employed to mitigate risks related to data privacy and security (Kaissis et al., 2020; Yang et al., 2019). Additionally, it is essential to ensure that data usage complies with relevant regulations, such as the EU's General Data Protection Regulation (GDPR) (General Data Protection Regulation, 2016).

7.1.2. Algorithmic bias and fairness

Some studies (Char et al., 2018; Obermeyer et al., 2019) suggest that AI systems may exhibit algorithmic bias, leading to inaccuracies in diagnosis for certain demographics (e.g., race, gender, or age groups), resulting in potential disparities in medical outcomes. For example, if a training dataset contains an insufficient number of samples from a specific group, the ARRG system may fail to diagnose diseases accurately in that population (Price & Cohen, 2019). To mitigate this issue, the following measures can be taken: Firstly, regular evaluation of models across different demographic groups to ensure that sensitivity and specificity are balanced. Secondly, adopting fairness-aware AI algorithms to reduce bias stemming from data imbalances (Char et al., 2018).

7.1.3. Transparency and explainability

AI models are often regarded as "black boxes", meaning that their decision-making processes lack explainability (Char et al., 2018; He et al., 2019; Jiang et al., 2017). This can decrease patient trust in ARRG-generated reports, especially when AI conclusions contradict those of

Table 9
Comparison of ARRG performance based on the IU X-ray dataset (Demner-Fushman et al., 2016). "T-B, R-B, A-M, M-F, K-I, R-L" denote "template-based models, retrieval-based models, attention mechanisms, multimodal fusion, knowledge integration, reinforcement learning", respectively. Additionally, "BL-1, BL-2, BL-3, BL-4, R-L, M, C, P, R, F-1" represent "BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR, CIDEr, and precision, recall, and F1-score (based on CheXpert Irvin et al., 2019)", respectively.

Methods	Visual extractor	Report generation	BL-1	BL-2	BL-3	BL-4	R-L	M	C	P	R	F-1
CNN-TRG (Pino et al., 2021)	DenseNet-121	T-B	0.273 (average of BLEU 1–4)				0.352	-	0.249	0.225	0.357	0.239
(Kale et al., 2023)	**-**	**T-B**	**0.775**	**0.699**	**0.658**	**0.627**	**0.817**	**0.782**	**0.853**	**-**	**-**	**-**
(Syeda-Mahmood et al., 2020)	VGG & ResNet-50	R-B	0.56	0.51	0.5	0.49	0.58	0.55	-	-	-	-
RTEX (Kougia et al., 2021)	DenseNet-121	R-B	0.55				0.202	-	-	0.193	0.222	-
(Akbar et al., 2023)	DenseNet-121	GRU	0.558	0.463	0.311	0.097	0.448	-	-	-	-	-
(Nguyen et al., 2021)	DenseNet-121	Transformer	0.515	0.378	0.293	0.235	0.436	0.219	-	-	-	-
(Sirshar et al., 2022)	VGG	LSTM + A-M	0.580	0.342	0.263	0.155	-	-	-	-	-	-
RadioBERT (Kaur, 2022)	**VGG-19**	**HLSTM** + Distill BERT	**0.772**	**0.770**	**0.768**	**0.767**	**0.897**	**-**	**0.556**	**-**	**-**	**-**
R2GenGPT (Wang et al., 2023c)	Swin Transformer	Llama2-7B	0.488	0.316	0.228	0.173	0.377	0.211	0.438	-	-	-
(Jing et al., 2017)	VGG-19	HLSTM + A-M	0.517	0.386	0.306	0.247	0.447	0.217	0.327	-	-	-
(Yuan et al., 2019)	ResNet-152	HLSTM + A-M	0.529	0.372	0.315	0.255	0.453	0.343	-	-	-	-
LIFMRG (Sun et al., 2024)	ViT	HLSTM + A-M	0.521	0.384	0.292	0.227	0.435	0.244	-	-	-	-
ASGMD (Xue et al., 2024)	ResNet-101	Transformer + A-M	0.489	0.326	0.232	0.173	0.397	0.206	-	-	-	-
DACG (Lang et al., 2025)	ResNet-101	Transformer + A-M	0.518	0.355	0.260	0.198	0.414	0.216	0.415	-	-	-
CheXPrune (Kaur, 2023)	VGG-19	HLSTM + A-M	0.543	0.445	0.374	0.320	0.598	-	0.322	-	-	-
(Wang et al., 2024d)	ResNet-101	Transformer + M-B	0.497	0.357	0.279	0.225	0.408	0.217	-	-	-	-
IFNet (Guo et al., 2024)	ResNet-152	Transformer + M-B	0.507	0.346	0.249	0.190	0.403	0.224	-	-	-	-
(Luan et al., 2023)	ResNet-50 + ViT	Transformer + M-B	0.540	0.424	0.362	0.322	0.479	0.246	-	-	-	-
(Cheddi et al., 2024)	MambaXray-VL-Large	BART + M-B	0.482	0.312	0.251	-	0.381	0.198	-	-	-	-
(Wang et al., 2024c)	Mamba	Llama2 + M-B	0.491	0.330	0.241	0.185	0.371	0.216	0.524	-	-	-
MATNet (Shang et al., 2022)	DenseNet-121	Transformer + A-M + M-B	0.518	0.387	0.308	0.254	0.446	0.222	-	-	-	-
DEKG (Wang et al., 2023a)	ResNet-101	Transformer + K-D	0.494	0.322	0.231	0.173	0.380	0.203	-	-	-	-
KiUT (Huang et al., 2023)	ResNet-101	Transformer + K-D	0.525	0.360	0.251	0.185	0.409	0.242	-	-	-	-
MKMIA (Zhao et al., 2023b)	ResNet-101	Transformer + K-D	0.513	0.340	0.245	0.188	0.399	0.216	-	-	-	-
(Zhang et al., 2022)	DenseNet-121	Transformer + K-D	0.505	0.379	0.303	0.251	0.446	0.218	-	-	-	-
DCL (Li et al., 2023a)	ViT	Transformer + K-D	-	-	-	0.163	0.383	0.193	0.586	-	-	-
KARGEN (Li et al., 2024b)	Swin Transformer	LLaMA2-7B	0.490	0.323	0.232	0.180	0.385	0.218	0.491	-	-	-
IVGN (Zheng et al., 2024)	ResNet-101	LSTM + A-M + K-D	0.523	0.357	0.258	0.191	0.405	0.226	-	-	-	-
CMM + RL (Qin & Song, 2022)	ResNet-101	Transformer + R-L	0.494	0.321	0.235	0.181	0.384	0.201	-	-	-	-
HReMRG-MR (Xu et al., 2023)	ResNet-101	LSTM + A-M + R-L	0.440	0.306	0.214	0.149	0.381	0.197	0.524	-	-	-
CADxReport (Kaur, 2022)	VGG-19	HLSTM + A-M + R-L	0.577	0.478	0.403	0.346	0.618	-	0.380	-	-	-
CLARA (Biswal et al., 2020)	DenseNet	Hybrid Model	0.512	0.402	0.281	0.254	-	-	0.425	-	-	-
MedWriter (Yang et al., 2021c)	DenseNet-121	Hybrid Model	0.471	0.336	0.238	0.166	0.382	-	0.345	-	-	-
MedNet (Nie & Liu, 2023)	CNN	Hybrid Model	0.491	0.349	0.255	0.198	0.426	-	0.383	-	-	-
RepsNet (Tanwani et al., 2022)	ResNeXt	Hybrid Model	0.58	0.44	0.32	0.27	-	-	-	-	-	-
TranSQ (Gao et al., 2024)	ViT	Hybrid Model	0.516	0.365	0.272	0.205	0.409	0.210	-	-	-	-

human radiologists, leaving patients uncertain about which assessment to believe. Possible optimizations include: Firstly, utilize visual explanation techniques. Methods like Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017) and MDNet (Medical Diagnosis Network) (Zhang et al., 2017) can provide insights into which areas the AI model focuses on during imaging analysis. However, because ARRG is a text-generation task rather than only a classification task, explanation should also cover how textual statements are produced. Cross-modal attention maps, evidence-linked sentence generation, and counterfactual analysis of generated findings can provide more task-relevant interpretability than heat maps alone. Secondly, develop multimodal ARRG models. Combining medical imaging, patient history, and laboratory test data may improve not only predictive performance but also the traceability of why particular report statements were generated.

7.1.4. Informed consent and patient autonomy

One of the four fundamental principles of medical ethics is patient autonomy. Physicians have an obligation to inform patients about whether their medical imaging data will be analysed by AI and ensure that patients have the right to decide whether they accept AI-assisted diagnosis (Gerke et al., 2020). Notably, ARRG is designed not to replace radiologists but to serve as an auxiliary tool. Human-centered care still

Table 10
Comparison of ARRG performance based on the MIMIC-CXR dataset (Johnson et al., 2019a,b). "T-B, R-B, A-M, M-F, K-I, R-L" denote "template-based models, retrieval-based models, attention mechanisms, multimodal fusion, knowledge integration, reinforcement learning", respectively. Additionally, "BL-1, BL-2, BL-3, BL-4, R-L, M, C, P, R, F-1" represent "BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR, CIDEr, and precision, recall, and F1-score (based on CheXpert Irvin et al., 2019)", respectively.

Methods	Visual extractor	Report generation	BL-1	BL-2	BL-3	BL-4	R-L	M	C	P	R	F-1
CNN-TRG (Pino et al., 2021)	DenseNet-121	T-B	0.094 (average of BLEU 1–4)				0.185	-	0.238	0.381	0.531	0.428
(Kale et al., 2023)	**-**	**T-B**	**0.833**	**0.807**	**0.749**	**0.785**	**0.833**	**0.861**	**0.861**	**-**	**-**	**-**
RTEX (Kougia et al., 2021)	DenseNet-121	R-B	0.59				0.205	-	-	0.229	0.284	-
(Nguyen et al., 2021)	DenseNet-121	Transformer	0.495	0.360	0.278	0.224	0.390	0.222	-	-	-	-
(Serra et al., 2023a)	Faster R-CNN	Transformer	0.486	0.366	0.295	0.246	0.423	0.216	-	0.553	0.597	0.516
RECAP (Hou et al., 2023a)	ViT	Transformer	0.415	0.257	0.171	0.119	0.285	0.164	-	0.381	0.443	0.391
HERGen (Wang et al., 2024a)	CvT	DistilGPT2	0.395	0.248	0.169	0.122	0.285	0.156	-	0.415	0.301	0.317
(Li et al., 2024a)	Faster R-CNN	GPT-4	0.395	0.260	0.178	0.131	0.261	0.161	-	0.469	0.470	0.441
R2GenGPT (Wang et al., 2023c)	Swin Transformer	Llama2-7B	0.411	0.267	0.186	0.134	0.297	0.160	0.269	0.392	0.387	0.389
(Sun et al., 2024)	ViT	HLSTM + A-M	0.428	0.274	0.211	0.155	0.327	0.201	-	0.662	0.440	0.528
ASGMD (Xue et al., 2024)	ResNet-101	Transformer + A-M	0.372	0.233	0.154	0.112	0.286	0.152	-	-	-	-
DACG (Lang et al., 2025)	ResNet-101	Transformer + A-M	0.398	0.249	0.167	0.117	0.290	0.162	-	0.422	0.405	0.389
(Mei et al., 2024)	DenseNet-121	Transformer + M-B	0.436	0.275	0.184	0.129	0.305	0.177	-	-	-	-
MambaXray -VL-Large (Wang et al., 2024c)	Mamba	Llama2 + M-B	0.422	0.268	0.184	0.133	0.289	0.167	0.241	-	-	-
NADM (Zhao et al., 2023a)	ResNet-50	Transformer + A-M + M-B	0.402	0.258	0.179	0.130	0.289	0.155	-	0.417	0.413	0.415
MATNet (Shang et al., 2022)	DenseNet-121	Transformer + A-M + M-B	0.506	0.370	0.288	0.233	0.395	0.221	-	0.454	0.391	0.405
DEKG (Wang et al., 2023a)	ResNet-101	Transformer + K-D	0.394	0.257	0.175	0.124	0.299	0.151	-	0.412	0.292	0.306
KiUT (Huang et al., 2023)	ResNet-101	Transformer + K-D	0.393	0.243	0.159	0.113	0.285	0.160	-	0.371	0.318	0.321
MKMIA (Zhao et al., 2023b)	ResNet-101	Transformer + K-D	0.399	0.242	0.158	0.109	0.275	0.152	-	0.482	0.505	0.493
(Zhang et al., 2022)	DenseNet-121	Transformer + K-D	0.491	0.358	0.278	0.225	0.389	0.215	-	-	-	-
DCL (Li et al., 2023a)	ViT	Transformer + K-D	-	-	-	0.109	0.284	0.150	0.281	-	-	-
KARGEN (Li et al., 2024b)	Swin Transformer	LLaMA2-7B	0.417	0.274	0.192	0.140	0.305	0.165	0.289	-	-	-
IVGN (Zheng et al., 2024)	ResNet-101	LSTM + A-M + K-D	0.377	0.236	0.158	0.112	0.280	0.144	-	0.415	0.422	0.398
CMM + RL (Qin & Song, 2022)	ResNet-101	Transformer + R-L	0.381	0.232	0.155	0.109	0.287	0.151	-	0.342	0.294	0.292
HReMRG-MR (Xu et al., 2023)	ResNet-101	LSTM + A-M + R-L	0.481	0.343	0.256	0.192	0.380	0.207	0.372	-	-	-
MedNet (Nie & Liu, 2023)	CNN	Hybrid Model	0.502	0.349	0.275	0.213	0.365	-	0.423	-	-	-
MedWriter (Yang et al., 2021c)	DenseNet-121	Hybrid Model	0.438	0.297	0.216	0.164	0.332	-	0.306	-	-	-
TranSQ (Gao et al., 2024)	ViT	Hybrid Model	0.423	0.261	0.171	0.116	0.286	0.168	-	0.482	0.563	0.519

relies on doctor-patient interaction to provide personalized treatment plans.

7.2. Doctors: responsibility and professional competency

From the perspective of doctors, the application of ARRG technology raises concerns about responsibility attribution and over-reliance leading to skill degradation.

7.2.1. Responsibility attribution and doctor-patient relationship

Although ARRG improves efficiency in radiology report generation, responsibility attribution remains a critical issue when errors or omissions occur (He et al., 2019). From a legal standpoint, radiologists must ultimately be responsible for the accuracy of reports, and ARRG cannot serve as the sole basis for medical decisions. Specifically, the following measures should be considered: Firstly, implementing human-in-the-loop (HITL) collaboration, where radiologists must review and sign off on AI-generated reports to ensure diagnostic accuracy. Secondly, establishing clear legal liability frameworks that define the responsibilities of doctors, hospitals, and ARRG system developers.

7.2.2. Over-reliance and skill degradation

Excessive reliance on ARRG systems may lead to the decline of radiologists' professional skills, negatively affecting long-term healthcare quality. As previously mentioned, a collaborative human-in-the-loop approach should be adopted to maximize strengths and minimize weaknesses. For instance: ARRG may be more sensitive than doctors in detecting small nodules, but AI can sometimes misidentify cross-sections of blood vessels as nodules, making radiologists' second confirmation essential. To address this issue, here are some recommendations. On the one hand, continuous training programs should be implemented to help radiologists understand the limitations of ARRG while maintaining their diagnostic proficiency. On the other hand, clinical experience must remain central in medical decision-making, with ARRG serving as an assistive tool rather than a substitute (Davenport & Kalakota, 2019; Topol, 2019).

7.3. Regulatory authorities: laws, standards, and ethical oversight

Different countries enforce varying medical regulations (Price & Cohen, 2019). For example, the EU's GDPR (General Data Protection Regulation, 2016) emphasizes data protection and patient privacy. ARRG systems must undergo rigorous clinical validation and comply with established international standards. Additionally, continuous monitoring and evaluation are essential to maintaining ARRG system reliability and ethical integrity.

ARRG technology holds great potential for enhancing radiologists' efficiency and improving medical services, but it also presents challenges concerning privacy protection, responsibility attribution, algorithmic fairness, and legal oversight. Moving forward, ARRG development should be guided by ethical and regulatory principles to ensure transparency and reliability while balancing the interests of patients, doctors, and regulatory authorities, ultimately realizing safe, fair, and efficient AI applications in healthcare (European Union, 2025; General Data Protection Regulation, 2016).

8. Limitations & future direction

Despite the increasing attention on ARRG and the progress achieved in this field, several limitations remain.

8.1. Radiology report formatting

A review of the literature reveals that very few studies focus on generating structured radiology reports, with most research emphasizing diagnostic text generation. The primary reason for this is the absence of a globally standardized radiology report format. Currently, radiology reports are predominantly categorized into free-text reports and structured reports. Given the lack of a unified reporting format within the radiology community, the datasets used to train ARRG models also lack consistency. Consequently, the generated reports from these models naturally follow no standardized format.

The absence of a globally unified radiology report format stems from the inherent advantages and limitations of both free-text and structured reporting systems. For instance, structured reports often lack the necessary flexibility to accurately characterize complex cases with nuanced lesion features. Conversely, structured reporting demonstrates superior performance in standardized scenarios such as tumor staging.

This dichotomy underscores the critical importance of developing ARRG systems that synergistically integrate the strengths of both free-text and structured formats - an approach that would significantly enhance clinical utility and workflow efficiency.

8.2. Medical experimental data

As highlighted in most ARRG-related review papers, current ARRG research primarily relies on 2D chest X-ray datasets. While some studies have begun incorporating 3D chest CT data, 3D lumbar spine MRI data, and 2D CT slices of other anatomical regions (e.g., brain scans), there is still a significant lack of data diversity.

Another major issue is the limited data scale, data scarcity, and accessibility challenges, which are discussed in detail in Section 4 and other review articles. Here, we introduce an additional concern: the absence of bone window imaging in datasets. Whether in 3D CT datasets (e.g., CT-RATE Dataset Hamamci et al., 2024a) or 2D CT slice datasets (e.g., CTRG-Brain-263K Dataset Tang et al., 2024), only soft tissue windows are provided, while bone window data is missing. Take a case from the CTRG-Brain-263K dataset (Tang et al., 2024) as an example (Fig. 8). The report states: "The bone window shows a linear low-density shadow in the left holoskeleton and skull base". However, the corresponding CT images provided only include soft tissue windows, with no bone window available. Since soft tissue windows alone do not clearly reveal bony structures, it is not possible to directly identify the "left temporal bone and skull base fracture".

For trauma patients, particularly those with cranial injuries, the absence of bone window imaging makes it difficult to determine the presence of skull fractures-an essential piece of clinical information. Due to this limitation, ARRG models trained on such datasets may fail to detect skull fractures or rib metastases, which are critical findings in radiology.

8.3. Methods

Currently, research on ARRG encompasses a wide range of methods, reflecting the fact that no single approach has yet emerged as a perfect solution. For example, in the study conducted by Kale et al., a template-based approach was adopted, and the experimental results demonstrated outstanding performance in NLG evaluation metrics (Kale et al., 2023). However, this does not imply that template-based methods are the optimal solution. This may be attributed to the fact that NLG evaluation metrics primarily stem from their reliance on n-gram matching and word order similarity to assess the resemblance between the generated and reference texts. Since template-based methods employ predefined sentence structures and fixed expressions, the generated reports exhibit a high degree of similarity to reference reports in terms of vocabulary selection, phrase order, and syntactic structure, thereby achieving higher scores in surface-level NLG metrics such as BLEU and ROUGE. Nevertheless, template-based methods lack flexibility and struggle to handle complex or rare cases.

In contrast, foundation models have demonstrated outstanding performance in image captioning tasks but have not performed as well as expected in ARRG tasks. This discrepancy may be primarily attributed to several factors. First, the size and quality of medical datasets are crucial factors affecting model performance. Compared to image captioning tasks, which typically rely on large-scale, high-quality annotated datasets, ARRG datasets are significantly smaller. This disparity in dataset scale may limit the generalization ability of foundation models. Second, there is a fundamental difference in task objectives. While image captioning focuses on describing visual content, ARRG must ensure the medical accuracy of the generated reports. That is, the text must not only be fluent but also adhere to clinical diagnostic standards. This task involves not only natural language generation but also requires the model to integrate medical knowledge for reasoning, ensuring the scientific validity and clinical applicability of the generated content. Moreover, there are notable differences between medical images and natural images. In medical imaging, lesions are typically much smaller than normal anatomical structures. This inherent data imbalance may impact model performance in ARRG tasks, making it challenging for models to accurately identify and describe abnormalities. Additionally, in real-world clinical scenarios, most medical images represent normal cases, while abnormal cases are relatively rare. This imbalance in class distribution further exacerbates the performance gap between ARRG and image captioning tasks, making it more difficult for models to generate high-quality textual descriptions in ARRG compared to image captioning. These factors collectively contribute to the inferior performance of foundation models in ARRG compared to their success in image captioning tasks.

Given that no single approach has proven to be universally optimal, future research should explore hybrid methods, which integrate the strengths of different approaches to enhance report accuracy and interpretability, such as combining template-based methods with foundation models. By employing Transformers or other foundation models, the sentence structures of templates can be dynamically adjusted to increase flexibility rather than being entirely fixed. Furthermore, considering the success of foundation models in image captioning, techniques such as fine-tuning, transfer learning, and pre-training can be explored to optimize ARRG. In addition, a fundamental distinction between ARRG and image captioning lies in the fact that ARRG is not merely about generating fluent text but also ensuring clinical accuracy, that is, whether

the generated content meets diagnostic standards. Therefore, future research should further explore knowledge integration methods to enhance the medical reliability of the reports, as well as multimodal methods that incorporate imaging features, patient history, and laboratory data to enable models to generate reports that are more aligned with real-world clinical environments.

8.4. Evaluation methods

The ultimate goal of ARRG research is to reduce the workload of radiologists by providing accurate and timely reports. In theory, human (expert) evaluation serves as the gold standard. However, a literature review reveals that human evaluations often lack objective metrics and instead rely on subjective assessments.

This raises a key issue: evaluation standards are highly dependent on the experience of the radiologists involved in the assessment process. Moreover, human evaluation is time-consuming, which ironically increases the workload for radiologists. Therefore, developing new evaluation methods is an urgent priority in ARRG research.

Another promising future direction is the development of LLM-assisted evaluation frameworks for automated radiology report generation. Recent studies in natural language generation suggest that strong large language models, when used with carefully designed prompts, scoring rubrics, or structured evaluation procedures, can achieve better alignment with human judgments than conventional surface-overlap metrics such as BLEU or ROUGE (Chiang & Lee, 2023; Kim et al., 2024a,b; Liu et al., 2023b). Related work further shows that rubric-guided, chain-of-thought-based, and evaluator-oriented language models may enable richer and more flexible assessments of generation quality than traditional automatic metrics alone, especially for dimensions such as coherence, consistency, and overall response quality (Kim et al., 2024a,b; Liu et al., 2023b). In addition, recent work on evaluator-oriented language models suggests a move toward more transparent, reproducible, and customizable evaluation systems beyond proprietary black-box models (Kim et al., 2024a,b).

For ARRG, this emerging direction is potentially relevant because clinically meaningful evaluation often requires assessment beyond lexical similarity alone. In particular, report evaluation may need to consider whether generated outputs preserve clinically important findings, maintain internal consistency across report sections, and avoid critical factual errors. However, LLM-based evaluation should still be interpreted cautiously. Existing studies have highlighted limitations such as prompt sensitivity, evaluator bias, and reproducibility concerns in LLM-based evaluation (Chiang & Lee, 2023; Kim et al., 2024a,b; Liu et al., 2023b). These limitations suggest that, in specialized domains such as radiology, general-purpose LLM evaluators may not always reliably capture task-specific clinical requirements. Therefore, in the near term, LLM-assisted evaluation is better viewed as a complementary component of hybrid evaluation frameworks, to be combined with structured factual metrics, report-specific clinical checklists, and targeted expert validation, rather than as a replacement for clinically grounded human assessment.

8.5. Ethical considerations

While ARRG technology holds great potential for enhancing radiologists' efficiency and optimizing healthcare services, it also introduces challenges related to privacy protection, accountability, algorithmic fairness, and legal regulations. In addition, future work should explicitly target ARRG-specific safety problems, including hallucination detection, faithful preservation of negation, omission of key abnormalities, alignment between Findings and Impression, correct handling of references to prior examinations, and safeguards against reports in which templated normal descriptions obscure rare or clinically important abnormalities.

Moving forward, the development of ARRG must be guided by ethical principles and regulatory frameworks to ensure transparency and reliability. Achieving a balance between patients, physicians, and regulatory bodies will be essential to realizing safe, fair, and efficient AI applications in healthcare.

9. Conclusion

This review examines the development of ARRG research from five key perspectives: report format, data, methodology, evaluation methods, and ethics. First, we show that not all cited ARRG datasets play the same role: some directly support image-to-report generation, whereas others mainly contribute auxiliary supervision, grounding, or evaluation signals. This distinction is important for interpreting both model design and reported benchmark performance. For example, in CT imaging, an ideal dataset should include both soft-tissue window and bone window images, yet most publicly available datasets currently only provide soft-tissue window data. Second, we examine ARRG methodologies through the complementary lenses of methodological paradigms and enhancement mechanisms. In this way, traditional template-based and retrieval-based approaches, encoder-decoder families, foundation models, and hybrid systems can be discussed alongside attention, multimodal fusion, knowledge integration, and reinforcement learning without conflating architectural levels. Third, we examine the clinical implications of report format. Free-text reports offer flexibility for complex or atypical cases, whereas structured reports improve standardization, completeness, and downstream data use. This supports the view that future ARRG systems may benefit from hybrid reporting targets that combine structured anchors with flexible natural-language description. Finally, we extend the ethical discussion by emphasizing ARRG-specific generative risks, including hallucination, omission of key abnormalities, negation errors, inconsistency between different report sections, incorrect references to prior examinations, and the risk that templated normal descriptions may obscure rare but clinically important abnormalities, in addition to broader concerns such as privacy, bias, and legal responsibility.

Given the rapid development of ARRG research, new datasets, models, and evaluation strategies may continue to emerge beyond the review window covered in this article (Wu et al., 2026). Looking ahead, progress in ARRG is likely to depend on a linked sequence of advances rather than isolated improvements. Better public and clinically richer datasets can support more realistic multimodal and knowledge integration modeling; stronger evaluation frameworks can then provide more reliable feedback on factual correctness, report structure, and clinical usefulness; and these foundations together are necessary for safer real-world deployment. In this context, future research should place particular emphasis on clinically grounded evaluation standards, including hybrid frameworks that combine structured factual metrics, report-specific clinical criteria, expert assessment, and cautiously used LLM-assisted evaluation. At the same time, continued attention to ethical, regulatory, and deployment-related challenges will remain essential for translating ARRG systems into trustworthy clinical practice.

CRediT authorship contribution statement

Lina Huang: Conceptualization, Methodology, Writing – original draft; **Tasin Islam:** Writing – review & editing; **Alina Miron:** Supervision, Writing – review & editing; **Kate Hone:** Supervision; **Yongmin Li:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data used in this review are from publicly available sources.

References

Abela, B., Abu-Khalaf, J., Yang, C.-W. R., Masek, M., & Gupta, A. (2022). Automated radiology report generation using a transformer-template system: Improved clinical accuracy and an assessment of clinical safety. In *Australasian joint conference on artificial intelligence* (pp. 530–543). Springer.

Achenbach, S., Fuchs, F., Goncalves, A., Kaiser-Albers, C., Ali, Z. A., Bengel, F. M., Dimmeler, S., Fayad, Z. A., Mebazaa, A., Meder, B. et al. (2022). Non-invasive imaging as the cornerstone of cardiovascular precision medicine. *European Heart Journal-Cardiovascular Imaging*, *23*(4), 465–475.

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D. et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, *5*(1), 4006.

Akbar, W., Haq, M. I. U., Soomro, A., Daudpota, S. M., Imran, A. S., & Ullah, M. (2023). Automated report generation: A GRU based method for chest x-rays. In *2023 4th international conference on computing, mathematics and engineering technologies (ICOMET)* (pp. 1–6). IEEE.

Al-Kafri, A. S., Sudirman, S., Hussain, A., Al-Jumeily, D., Natalia, F., Meidia, H., Afriliana, N., Al-Rashdan, W., Bashtawi, M., & Al-Jumaily, M. (2019). Boundary delineation of MRI images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks. *IEEE Access*, *7*, 43487–43501.

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, *24*, 100557.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Alsharid, M., Sharma, H., Drukker, L., Chatelain, P., Papageorghiou, A. T., & Noble, J. A. (2019). Captioning ultrasound images automatically. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd International conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22* (pp. 338–346). Springer.

Aronson, A. R., & Lang, F.-M. (2010). An overview of metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

Babar, Z., van Laarhoven, T., & Marchiori, E. (2021a). Encoder-decoder models for chest x-ray report generation perform no better than unconditioned baselines. *Plos One*, *16*(11), e0259639.

Babar, Z., van Laarhoven, T., Zanzotto, F. M., & Marchiori, E. (2021b). Evaluating diagnostic content of AI-generated radiology reports of chest x-rays. *Artificial Intelligence in Medicine*, *116*, 102075.

Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024). M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578.

Baidu, P., Online resource. https://pan.baidu.com/s/1Rm-0-a7jWJYcbpj3-Zo71A. Accessed: 18 January 2025.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H. et al. (2024). Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449.

Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., de Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Schwaighofer, A. et al. (2023a). MS-CXR-T: Learning to exploit temporal structure for biomedical vision-language processing. PhysioNet. https://doi.org/10.13026/pg10-j984. https://physionet.org/content/ms-cxr-t/1.0.0/.Version 1.0.0. RRID:SCR_007345.

Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A. et al. (2023b). Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15016–15027).

Bécares-Martínez, C., López-Llames, A., Martín-Pagán, A., Cores-Prieto, A. E., Arroyo-Domingo, M., Marco-Algarra, J., & Morales-Suárez-Varela, M. (2020). Cervical spine radiographs in patients with vertigo and dizziness. *La Radiologia Medica*, *125*, 272–279.

Beddiar, D.-R., Oussalah, M., & Seppänen, T. (2023). Automatic captioning for medical imaging (MIC): A rapid review of literature. *Artificial Intelligence Review*, *56*(5), 4019–4076.

Benger, J. R., & Lyburn, I. D. (2003). What is the effect of reporting all emergency department radiographs? *Emergency Medicine Journal*, *20*(1), 40–43.

Bergomi, L., Buonocore, T. M., Antonazzo, P., Alberghi, L., Bellazzi, R., Preda, L., Bortolotto, C., & Parimbelli, E. (2024). Reshaping free-text radiology notes into structured reports with generative question answering transformers. *Artificial Intelligence in Medicine*, *154*, 102924.

Bernardi, M. L., & Cimitile, M. (2024). Report generation from x-ray imaging by retrieval-augmented generation and improved image-text matching. In *2024 International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Bilic, P., Christ, P., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G. E. H., Chartrand, G. et al. (2023). The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, *84*, 102680.

BIMCV. Padchest: A large chest x-ray image dataset. http://bimcv.cipf.es/bimcv-projects/padchest/. Accessed: 18 January 2025.

Biswal, S., Xiao, C., Glass, L. M., Westover, B., & Sun, J. (2020). Clara: Clinical report auto-completion. In *Proceedings of the web conference 2020* (pp. 541–550).

Boag, W., Hsu, T.-M. H., McDermott, M., Berner, G., Alesentzer, E., & Szolovits, P. (2020). Baselines for chest x-ray report generation. In *Machine learning for health workshop* (pp. 126–140). PMLR.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267–D270.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J. et al. (2022). Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision* (pp. 1–21). Springer.

Brook, O. R., Brook, A., Vollmer, C. M., Kent, T. S., Sanchez, N., & Pedrosa, I. (2015). Structured reporting of multiphasic CT for pancreatic cancer: Potential effect on staging and surgical planning. *Radiology*, *274*(2), 464–472.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bustos, A., Pertusa, A., Salinas, J.-M., & De La Iglesia-Vaya, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, *66*, 101797.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., & McDonald, C. J. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, *33*(2), 577–590.

Caranci, F., Leone, G., Ponsiglione, A., Muto, M., Tortora, F., Muto, M., Cirillo, S., Brunese, L., & Cerase, A. (2020). Imaging findings in hypophysitis: A review. *La Radiología Médica*, *125*, 319–328.

Castro, D. C., Bustos, A., Bannur, S., Hyland, S. L., Bouzid, K., Wetscherek, M. T., Sánchez-Valverde, M. D., Jaques-Pérez, L., Pérez-Rodríguez, L., Takeda, K. et al. (2024). Padchest-GR: A bilingual chest x-ray dataset for grounded radiology report generation. arXiv preprint arXiv:2411.05085.

Chambon, P., Delbrouck, J.-B., Sounack, T., Huang, S.-C., Chen, Z., Varma, M., Truong, S. Q. H., Chuong, C. T., & Langlotz, C. P. (2024). Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. https://arxiv.org/abs/2405.19538.

Chand, R. B., Thapa, N., Paudel, S., Pokharel, G. B., Joshi, B. R., & Pant, D. K. (2013). Evaluation of image quality in chest radiographs. *Journal of Institute of Medicine Nepal*, *35*(1), 50–52.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care-addressing ethical challenges. *New England Journal of Medicine*, *378*(11), 981–983.

Charalampakos, F., Karatzas, V., Kougia, V., Pavlopoulos, J., & Androutsopoulos, I. (2021). AUEB NLP group at imageclefmed caption tasks 2021. In *CLEF (working notes)* (pp. 1184–1200).

Cheddi, F., Habbani, A., & Nait-Charif, H. (2024). A multi-modal feature fusion-based approach for chest x-ray report generation. In *2024 11th international conference on wireless networks and mobile communications (WINCOM)* (pp. 1–7). IEEE.

Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., Guo, L., Han, J., Liu, T., Liu, J. et al. (2024a). 3d-CT-GPT: Generating 3D radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330.

Chen, Z., Bie, Y., Jin, H., & Chen, H. (2024b). Large language model with region-guided referring and grounding for CT report generation. arXiv preprint arXiv:2411.15539.

Chen, Z., Shen, Y., Song, Y., & Wan, X. (2022). Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258.

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.

Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J. M. J., Youssef, A., Cohen, J. P., Reis, E. P. et al. (2024c). Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208.

Chiang, C.-H., & Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? https://arxiv.org/abs/2305.01937.

Chng, S. Y., Tern, P. J. W., Kan, M. R. X., & Cheng, L. T. E. (2023). Automated labelling of radiology reports using natural language processing: Comparison of traditional and newer methods. *Health Care Science*, *2*(2), 120–128.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In *International conference on machine learning* (pp. 2067–2075). PMLR.

Codalab Competitions (2017). Liver tumor segmentation challenge. https://competitions.codalab.org/competitions/17094. Accessed: 20 January 2025.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10578–10587).

Cramer, J. A., Eisenmenger, L. B., Pierson, N. S., Dhatt, H. S., Heilbrun, M. E. et al. (2014). Structured and templated reporting: An overview. *Applied Radiology*, *43*(8), 18–21.

Dalla Serra, F., Clackett, W., MacKinnon, H., Wang, C., Deligianni, F., Dalton, J., & O'Neil, A. Q. (2022). Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing (volume 1: long papers)* (pp. 615–624).

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98.

Dejl, A., Khanna, S., Pile, P. T., Yoon, K., Truong, S. Q. H., Duong, H., Saenz, A., & Rajpurkar, P. (2024). Radgraph2: Tracking findings over time in radiology reports. *Phy-*

sioNet. Version 1.0.0 https://doi.org/10.13026/q65y-9688
Delbrouck, J.-B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., & Langlotz, C. P. (2022). Improving the factual correctness of radiology report generation with semantic rewards. arXiv preprint arXiv:2210.12186.
Delbrouck, J.-B., Xu, J., Moll, J., Thomas, A., Chen, Z., Ostmeier, S., Azhar, A., Li, K. Z., Johnston, A., Bluethgen, C. et al. (2025). Automated structured radiology report generation. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 26813–26829).
Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304–310.
Deng, X., He, X., Zhou, Y., Cai, S., Cai, C., & Chen, Z. (2024). MvkeTR: Chest CT report generation with multi-view perception and knowledge enhancement. arXiv preprint arXiv:2411.18309.
Di Piazza, T. (2024). Ct-agrg: Automated abnormality-guided report generation from 3D chest CT volumes. arXiv preprint arXiv:2408.11965.
Dick, J., Darras, K. E., Lexa, F. J., Denton, E., Ehara, S., Galloway, H., Jankharia, B., Kassing, P., Kumamaru, K. K., Mildenberger, P. et al. (2021). An international survey of quality and safety programs in radiology. *Canadian Association of Radiologists Journal*, *72*(1), 135–141.
Dimarco, M., Cannella, R., Pellegrino, S., Iadicola, D., Tutino, R., Allegra, F., Castiglione, D., Salvaggio, G., Midiri, M., Brancatelli, G. et al. (2020). Impact of structured report on the quality of preoperative CT staging of pancreatic ductal adenocarcinoma: Assessment of intra-and inter-reader variability. *Abdominal Radiology*, *45*, 437–448.
Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).
Dos Santos, D. P., Hempel, J.-M., Mildenberger, P., Klöckner, R., & Persigehl, T. (2019). Structured reporting in clinical routine. In *Röfo-fortschritte auf dem gebiet der röntgenstrahlen und der bildgebenden verfahren* (pp. 33–39). © Georg Thieme Verlag KG (*vol. 191*).
Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
D'Orsi, C. J., Sickles, E. A., Mendelson, E. B., & Morris, E., et al. (2013). ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology.
Eghtedari, M., Chong, A., Rakow-Penner, R., & Ojeda-Fournier, H. (2021). Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. *American Journal of Roentgenology*, *216*(4), 860–873.
Ehab, W., Huang, L., & Li, Y. (2024). Unet and variants for medical image segmentation. *International Journal of Network Dynamics and Intelligence*, *3*, 1–24. https://bura.brunel.ac.uk/handle/2438/28394.
Elyan, E., Vuttipittayamongkol, P., Johnston, P., Martin, K., McPherson, K., Moreno-García, C. F., Jayne, C., & Sarker, M. M. K. (2022). Computer vision and machine learning for medical image analysis: Recent advances, challenges, and way forward. *Artificial Intelligence Surgery*, *2*(1), 24–45.
Emaminejad, N., Qian, W., Guan, Y., Tan, M., Qiu, Y., Liu, H., & Zheng, B. (2015). Fusion of quantitative image and genomic biomarkers to improve prognosis assessment of early stage lung cancer patients. *IEEE Transactions on Biomedical Engineering*, *63*(5), 1034–1043.
European Union (2025). Eu AI act. Accessed: 06 March 2025 https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
European Society of Radiology (ESR) (2011). Good practice for radiological reporting. Guidelines from the European society of radiology (ESR). *Insights into Imaging*, *2*(2), 93–96.
European Society of Radiology (ESR) (2018). ESR paper on structured reporting in radiology. *Insights into Imaging*, *9*, 1–7.
European Society of Radiology (ESR) (2023). ESR paper on structured reporting in radiology-update 2023. *Insights into Imaging*, *14*(1), 199.
Faggioni, L., Coppola, F., Ferrari, R., Neri, E., & Regge, D. (2017). Usage of structured reporting in radiological practice: Results from an Italian online survey. *European Radiology*, *27*, 1934–1943.
FB, R. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, *51*, 114–116.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.
Flusberg, M., Ganeles, J., Ekinci, T., Goldberg-Stein, S., Paroder, V., Kobi, M., & Chernyak, V. (2017). Impact of a structured report template on the quality of CT and MRI reports for hepatocellular carcinoma diagnosis. *Journal of the American College of Radiology*, *14*(9), 1206–1211.
Gajbhiye, G. O., Nandedkar, A. V., & Faye, I. (2022). Translating medical image to radiological report: Adaptive multilevel multi-attention approach. *Computer Methods and Programs in Biomedicine*, *221*, 106853.
Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L. J., & Bradley, A. P. (2019). Producing radiologist-quality reports for interpretable deep learning. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)* (pp. 1275–1279). IEEE.
Ganeshan, D., Duong, P.-A. T., Probyn, L., Lenchik, L., McArthur, T. A., Retrouvey, M., Ghobadi, E. H., Desouches, S. L., Pastel, D., & Francis, I. R. (2018). Structured reporting in radiology. *Academic Radiology*, *25*(1), 66–73.
Gao, D., Kong, M., Zhao, Y., Huang, J., Huang, Z., Kuang, K., Wu, F., & Zhu, Q. (2024). Simulating doctors' thinking logic for chest x-ray report generation via transformer-based semantic query learning. *Medical Image Analysis*, *91*, 102982.
Gatt, M. E., Spectre, G., Paltiel, O., Hiller, N., & Stalnikowicz, R. (2003). Chest radiographs in the emergency department: Is the radiologist really necessary? *Postgraduate Medical Journal*, *79*(930), 214–217.
General Data Protection Regulation (2016). General Data Protection Regulation. Accessed: 06 March 2025 https://gdpr-info.eu/.
Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–336). Elsevier.
Goel, A. K., DiLella, D., Dotsikas, G., Hilts, M., Kwan, D., & Paxton, L. (2019). Unlocking radiology reporting data: An implementation of synoptic radiology reporting in low-dose CT cancer screening. *Journal of Digital Imaging*, *32*, 1044–1051.
Granata, V., Faggioni, L., Grassi, R., Fusco, R., Reginelli, A., Rega, D., Maggialetti, N., Buccicardi, D., Frittoli, B., Rengo, M. et al. (2022). Structured reporting of computed tomography in the staging of colon cancer: A delphi consensus proposal. *La Radiologia Medica*, *127* (1), 21–29.
Guo, Y., Hou, X., Liu, Z., & Zhang, Y. (2024). Ifnet: An image-enhanced cross-modal fusion network for radiology report generation. In *International symposium on bioinformatics research and applications* (pp. 286–297). Springer.
Hamamci, I. E., Er, S., Almas, F., Simsek, A. G., Esirgun, S. N., Dogan, I., Dasdelen, M. F., Durugol, O. F., Wittmann, B., Amiranashvili, T. et al. (2024a). Developing generalist foundation models from a multimodal dataset for 3D computed tomography. *Research Square*. https://doi.org/10.21203/rs.3.rs-5271327/v1
Hamamci, I. E., Er, S., Almas, F., Simsek, A. G., Esirgun, S. N., Dogan, I., Dasdelen, M. F., Wittmann, B., Simsar, E., Simsar, M. et al. (2024b). A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv preprint arXiv:2403.17834.
Hamamci, I. E., Er, S., & Menze, B. (2024c). Ct2rep: Automated radiology report generation for 3d medical imaging. In *International conference on medical image computing and computer-assisted intervention* (pp. 476–486). Springer.
Hamamci, I. E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A. G., Esirgun, S. N., Almas, F., Doğan, I., Dasdelen, M. F. et al. (2025). Generatect: Text-conditional generation of 3D chest ct volumes. In *European conference on computer vision* (pp. 126–143). Springer.
Han, Z., Wei, B., Leung, S., Chung, J., & Li, S. (2018). Towards automatic report generation in spine radiology using weakly supervised framework. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11* (pp. 185–193). Springer.
Han, Z., Wei, B., Xi, X., Chen, B., Yin, Y., & Li, S. (2021). Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical Image Analysis*, *67*, 101872.
Haroun, R. R., Al-Hihi, M. M., & Abujudeh, H. H. (2019). The pros and cons of structured reports. *Current Radiology Reports*, *7*, 1–4.
Harzig, P., Chen, Y.-Y., Chen, F., & Lienhart, R. (2019). Addressing data bias problems for chest x-ray image report generation. arXiv preprint arXiv:1908.02123.
He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, P. (2001). The digital database for screening mammography, IWDM-2000. In *Fifth international workshop on digital mammography., medical physics publishing* (pp. 212–218).
Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, *32*, 11135–11145.
Hochreiter, S. (1997). Long short-term memory. *Neural computation* MIT-Press.
Hou, W., Cheng, Y., Xu, K., Hu, Y., Li, W., & Liu, J. (2024). Icon: Improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. arXiv preprint arXiv:2402.12844.
Hou, W., Cheng, Y., Xu, K., Li, W., & Liu, J. (2023a). Recap: Towards precise radiology report generation via dynamic disease progression reasoning. arXiv preprint arXiv:2310.13864.
Hou, W., Xu, K., Cheng, Y., Li, W., & Liu, J. (2023b). Organ: Observation-guided radiology report generation via tree reasoning. arXiv preprint arXiv:2306.06466.
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
Huang, L., Miron, A., Hone, K., & Li, Y. (2024). Segmenting medical images: From UNet to res-UNet and nnUNet. In *2024 IEEE 37th international symposium on computer-based medical systems (CBMS)* (pp. 483–489). IEEE.
Huang, X., Yan, F., Xu, W., & Li, M. (2019). Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, *7*, 154808–154817.
Huang, Z., Zhang, X., & Zhang, S. (2023). Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19809–19818).
Ierardi, A. M., Wood, B. J., Arrichiello, A., Bottino, N., Bracchi, L., Forzenigo, L., Andrisani, M. C., Vespro, V., Bonelli, C., Amalou, A. et al. (2020). Preparation of a radiology department in an Italian hospital dedicated to COVID-19 patients. *La Radiologia Medica*, *125*, 894–901.
Ionescu, B., Müller, H., Drăgulinescu, A. M., Popescu, A., Idrissi-Yaghir, A., García Seco de Herrera, A., Andrei, A., Stan, A., Storås, A. M., Abacha, A. B. et al. (2023). ImageCLEF 2023 highlight: Multimedia retrieval in medical, social media and content recommendation applications. In *European conference on information retrieval* (pp. 557–567). Springer.
Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. et al. (2019). Chexpert: A large chest radiograph dataset

with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 590–597). (*vol. 33*).
Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S. et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, *33*(2), 233–245.
Jain, S., Agrawal, A., Saporta, A., Truong, S. Q. H., Duong, D. N., Bui, T., Chambon, P., Zhang, Y., Lungren, M. P., Ng, A. Y. et al. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463.
Javaid, M., Haleem, A., Singh, R. P., & Ahmed, M. (2024). Computer vision to enhance healthcare domain: An overview of features, implementation, and opportunities. *Intelligent Pharmacy*, *2* (6), 792–803.
Jeong, J., Tian, K., Li, A., Hartung, S., Adithan, S., Behzadi, F., Calle, J., Osayande, D., Pohlen, M., & Rajpurkar, P. (2024). Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical imaging with deep learning* (pp. 978–990). PMLR.
Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101.
Jing, B., Wang, Z., & Xing, E. (2020). Show, describe and conclude: On exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274.
Jing, B., Xie, P., & Xing, E. (2017). On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195.
Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., & Horng, S. (2019a). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317.
Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019b). Mimic-CXR-JPG, A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.
Johnson, A. J., Chen, M. Y. M., Zapadka, M. E., Lyders, E. M., & Littenberg, B. (2010). Radiology report clarity: A cohort study of structured reporting compared with conventional dictation. *Journal of the American College of Radiology*, *7*(7), 501–506.
Jorg, T., Kämpgen, B., Feiler, D., Müller, L., Düber, C., Mildenberger, P., & Jungmann, F. (2023). Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing. *Insights into Imaging*, *14*(1), 47.
Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305–311.
Kale, K., Jadhav, K. et al. (2023). Replace and report: NLP assisted radiology report generation. arXiv preprint arXiv:2306.17180.
Kaur, N., & Mittal, A. (2022a). Cadxreport: Chest x-ray report generation using co-attention mechanism and reinforcement learning. *Computers in Biology and Medicine*, *145*, 105498.
Kaur, N., & Mittal, A. (2022b). RadioBERT: A deep learning-based system for medical report generation from chest x-ray images using contextual embeddings. *Journal of Biomedical Informatics*, *135*, 104220.
Kaur, N., & Mittal, A. (2023). CheXPrune: Sparse chest x-ray report generation model using multi-attention and one-shot global pruning. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7485–7497.
Kaur, N., Mittal, A., & Singh, G. (2022). Methods for automatic generation of radiological reports of chest radiographs: A comprehensive survey. *Multimedia Tools and Applications*, *81*(10), 13409–13439.
Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 2). Minneapolis, Minnesota (*Vol. 1*).
Khanna, S., Dejl, A., Yoon, K., Truong, S. Q. H., Duong, H., Saenz, A., & Rajpurkar, P. (2023). Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine learning for healthcare conference* (pp. 381–402). PMLR.
Khurana, A., Nelson, L. W., Myers, C. B., Akisik, F., Jeffrey, B. R., Miller, F. H., Mittal, P., Morgan, D., Mortele, K., Poullos, P. et al. (2020). Reporting of acute pancreatitis by radiologists-time for a systematic change with structured reporting template. *Abdominal Radiology*, *45*, 1277–1289.
Kim, S., Nooralahzadeh, F., Rohanian, M., Fujimoto, K., Nishio, M., Sakamoto, R., Rinaldi, F., & Krauthammer, M. (2023). Boosting radiology report generation by infusing comparison prior. arXiv preprint arXiv:2305.04561.
Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., & Seo, M. (2024a). Prometheus: Inducing fine-grained evaluation capability in language models. https://arxiv.org/abs/2310.08491.
Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., & Seo, M. (2024b). Prometheus 2: An open source language model specialized in evaluating other language models. https://arxiv.org/abs/2405.01535.
Kisilev, P., Sason, E., Barkan, E., & Hashoul, S. (2016). Medical image description using multi-task-loss CNN. In *Deep learning and data labeling for medical applications: First international workshop, LABELS 2016, and second international workshop, DLMIA 2016, held in conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1* (pp. 121–129). Springer.
Kougia, V., Pavlopoulos, J., Papapetrou, P., & Gordon, M. (2021). Rtex: A novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams. *Journal of the American Medical Informatics Association*, *28*(8), 1651–1659.
Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 317–325).
Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*, 32–73.
Lang, W., Liu, Z., & Zhang, Y. (2025). Dacg: Dual attention and context guidance model for radiology report generation. *Medical Image Analysis*, *99*, 103377.
Langlotz, C. P. (2006). Radlex: A new method for indexing online educational materials. *RadioGraphics*, *26*(6), 1595–1597. https://doi.org/10.1148/rg.266065168.
Lecchi, M., Fossati, P., Elisei, F., Orecchia, R., & Lucignani, G. (2008). Current concepts on imaging in radiotherapy. *European Journal of Nuclear Medicine and Molecular Imaging*, *35*, 821–837.
Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019a). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6666–6673). (*Vol. 33*).
Li, H., Wang, H., Sun, X., He, H., & Feng, J. (2024a). Prompt-guided generation of structured chest x-ray report using a pre-trained LLM. In *2024 IEEE International conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.
Li, J., Li, S., Hu, Y., & Tao, H. (2022). A self-guided framework for radiology report generation. In *International conference on medical image computing and computer-assisted intervention* (pp. 588–598). Springer.
Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., & Chang, X. (2023a). Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3334–3343).
Li, M., Liu, R., Wang, F., Chang, X., & Liang, X. (2023b). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, *26*(1), 253–270.
Li, X., Cao, R., & Zhu, D. (2019b). Vispi: Automatic visual perception and interpretation of chest x-rays. arXiv preprint arXiv:1906.05190.
Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in Neural Information Processing Systems*, *31*.
Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., & Zhou, L. (2023c). A comprehensive study of GPT-4v's multimodal capabilities in medical imaging. *medRxiv*, https://doi.org/10.1101/2023.11.03.23298067.
Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., & Zhou, L. (2025). S-RRG-Bench: Structured radiology report generation with fine-grained evaluation framework. *Meta-Radiology*, *3*(4), 100171. https://doi.org/10.1016/j.metrad.2025.100171.
Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., & Zhou, L. (2024b). Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International conference on medical image computing and computer-assisted intervention* (pp. 382–392). Springer.
Liang, H., Jiang, M., Liang, R., & Zhao, Q. (2018). Capvis: Toward better understanding of visual-verbal saliency consistency. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(1), 1–23.
Liao, Y., Liu, H., & Spasić, I. (2023). Deep learning approaches to automatic radiology report generation: A systematic review. *Informatics in Medicine Unlocked*, *39*, 101273.
PEIR Digital Library (2025). https://peir.path.uab.edu/library/index.php?/category/106. Accessed: 19 January 2025.
Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).
Lin, M. (2013). Network in network. arXiv preprint arXiv:1312.4400.
Liu, C., Tian, Y., & Song, Y. (2023a). A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199.
Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021a). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13753–13762).
Liu, F., You, C., Wu, X., Ge, S., Sun, X. et al. (2021b). Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, *34*, 16266–16279.
Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. In *Machine learning for healthcare conference* (pp. 249–269). PMLR.
Liu, G., Liao, Y., Wang, F., Zhang, B., Zhang, L., Liang, X., Wan, X., Li, S., Li, Z., Zhang, S. et al. (2021c). Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(9), 3786–3797.
Liu, K., Ma, Z., Xie, K., Jiao, Z., & Miao, Q. (2024). MCL: Multi-view enhanced contrastive learning for chest x-ray report generation. arXiv preprint arXiv:2411.10224.
Liu, X., Xin, J., Shen, Q., Huang, Z., & Wang, Z. (2025). Automatic medical report generation based on deep learning: A state of the art survey. *Computerized Medical Imaging and Graphics*, *120*, 102486. https://doi.org/10.1016/j.compmedimag.2024.102486.
Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023b). G-EVAL: NLG evaluation using GPT-4 with better human alignment. https://arxiv.org/abs/2303.16634.
Lovelace, J., & Mortazavi, B. (2020). Learning to generate clinically coherent chest x-ray reports. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1235–1243).
Loveymi, S., Dezfoulian, M. H., & Mansoorizadeh, M. (2020). Generate structured radiology report from CT images using image annotation techniques: Preliminary results with liver CT. *Journal of Digital Imaging*, *33*(2), 375–390.
Loveymi, S., Dezfoulian, M. H., & Mansoorizadeh, M. (2021). Automatic generation of structured radiology reports for volumetric computed tomography images using question-specific deep feature extraction and learning. *Journal Of Medical Signals & Sensors*, *11*(3), 194–207.
Luan, Q., Pan, H., Zhang, K., Shi, K., & Jia, X. (2023). Enriching semantic features for medical report generation. In *Ccf international conference on natural language processing and chinese computing* (pp. 469–480). Springer.
Marcal, L. P., Fox, P. S., Evans, D. B., Fleming, J. B., Varadhachary, G. R., Katz, M. H., & Tamm, E. P. (2015). Analysis of free-form radiology dictations for completeness and clarity for pancreatic cancer staging. *Abdominal Imaging*, *40*, 2391–2397.

Marrie, T. J. (1997). Survey of physicians concerning the use of chest radiography in the diagnosis of pneumonia in out-patients. *Canadian Journal of Infectious Diseases and Medical Microbiology*, *8*(2), 95–98.
Marvasti, N. B., del Mar Roldan, G. M., Üsküdarli, S., Montes, J. F. A., & Acar, B. (2015). Overview of the imageCLEF 2015 liver CT annotation task. CLEF Working Notes.
Mazzone, P. J., Silvestri, G. A., Patel, S., Kanne, J. P., Kinsinger, L. S., Wiener, R. S., Hoo, G. S., & Detterbeck, F. C. (2018). Screening for lung cancer: CHEST guideline and expert panel report. *Chest*, *153*(4), 954–985.
Medicine, S. (2024). Chexpert plus. https://aimi.stanford.edu/datasets/chexpert-plus. Dataset accessed: 11 April 2026. https://doi.org/10.71718/6nvz-pm34
Mei, X., Mao, R., Cai, X., Yang, L., & Cambria, E. (2024). Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 4699–4708).
Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andía, M., Tejos, C., Prieto, C., & Capurro, D. (2022). A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–40.
Mondal, C., Pham, D.-S., Tan, T., Gedeon, T., & Gupta, A. (2023). Transformers are all you need to generate automatic report from chest x-ray images. In *2023 International conference on digital image computing: Techniques and applications (DICTA)* (pp. 387–394). IEEE.
Monshi, M. M. A., Poon, J., & Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, *106*, 101878.
Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, *19*(2), 236–248.
Morgan, T. A., Helibrun, M. E., & Kahn, C. E., Jr (2014). Reporting initiative of the radiological society of North America: Progress and new directions. *Radiology*, *273*(3), 642–645. https://doi.org/10.1148/radiol.14141227.
Mork, J. G., Jimeno-Yepes, A., Aronson, A. R. et al. (2013). The NLM medical text indexer system for indexing biomedical literature. *CEUR Workshop Proceedings*, *1094*, 1–6.
Neri, E., Coppola, F., Larici, A. R., Sverzellati, N., Mazzei, M. A., Sacco, P., Dalpiaz, G., Feragalli, B., Miele, V., & Grassi, R. (2020). Structured reporting of chest CT in COVID-19 pneumonia: A consensus proposal. *Insights into Imaging*, *11*, 1–9.
Neri, E., Granata, V., Montemezzi, S., Belli, P., Bernardi, D., Brancato, B., Caumo, F., Calabrese, M., Coppola, F., Cossu, E. et al. (2022). Structured reporting of x-ray mammography in the first diagnosis of breast cancer: A delphi consensus proposal. *La Radiologia Medica*, *127*(5), 471–483.
Nguyen, D., Chen, C., He, H., & Tan, C. (2023). Pragmatic radiology report generation. In *Machine learning for health (ML4h)* (pp. 385–402). PMLR.
Nguyen, H. T. N., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., & Cheng, L. (2021). Automated generation of accurate\& fluent medical x-ray reports. arXiv preprint arXiv:2108.12126.
Ni, J., Hsu, C.-N., Gentili, A., & McAuley, J. (2020). Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. arXiv preprint arXiv:2010.02467.
Nicolson, A., Dowling, J., & Koopman, B. (2023). Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, *144*, 102633.
Nie, P., & Liu, X. (2023). Mednet: A dual-copy mechanism for medical report generation from images. In *International conference on artificial neural networks* (pp. 469–481). Springer.
Nishino, T., Miura, Y., Taniguchi, T., Ohkuma, T., Suzuki, Y., Kido, S., & Tomiyama, N. (2022). Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7123–7138).
Nobel, J. M., Kok, E. M., & Robben, S. G. F. (2020). Redefining the structure of structured reporting in radiology. *Insights into Imaging*, *11*(1), 10.
Nooralahzadeh, F., Gonzalez, N. P., Frauenfelder, T., Fujimoto, K., & Krauthammer, M. (2021). Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777.
Nörenberg, D., Sommer, W. H., Thasler, W., D'Haese, J., Rentsch, M., Kolben, T., Schreyer, A., Rist, C., Reiser, M., & Armbruster, M. (2017). Structured reporting of rectal magnetic resonance imaging in suspected primary rectal cancer: Potential benefits for surgical planning and interdisciplinary communication. *Investigative Radiology*, *52*(4), 232–239.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.
Olthof, A. W., Leusveld, A. L. M., de Groot, J. C., Callenbach, P. M. C., & van Ooijen, P. M. A. (2020). Contextual structured reporting in radiology: Implementation and long-term evaluation in improving the communication of critical findings. *Journal of Medical Systems*, *44*(9), 148.
Pang, T., Li, P., & Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, *22*(1), 48.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).
Park, J., Kim, S., Yoon, B., Hyun, J., & Choi, K. (2024). M4CXR: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. arXiv preprint arXiv:2408.16213.
Park, J., Pillarisetty, V. G., Brennan, M. F., Jarnagin, W. R., D'Angelica, M. I., DeMatteo, R. P., Coit, D. G., Janakos, M., & Allen, P. J. (2010). Electronic synoptic operative reporting: Assessing the reliability and completeness of synoptic reports for pancreatic resection. *Journal of the American College of Surgeons*, *211*(3), 308–315.
Pelka, O., Koitka, S., Rückert, J., Nensa, F., & Friedrich, C. M. (2018). Radiology objects in context (roco): A multimodal image dataset. In *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis: 7th joint international workshop, CVII-STENT 2018 and third international workshop, LABELS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3* (pp. 180–189). Springer.
Pellegrini, C., Keicher, M., Özsoy, E., & Navab, N. (2023). Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International conference on medical image computing and computer-assisted intervention* (pp. 409–419). Springer.
Pesapane, F., Tantrige, P., De Marco, P., Carriero, S., Zugni, F., Nicosia, L., Bozzini, A. C., Rotili, A., Latronico, A., Abbate, F. et al. (2023). Advancements in standardizing radiological reports: A comprehensive review. *Medicina*, *59*(9), 1679.
Pino, P., Parra, D., Besa, C., & Lagos, C. (2021). Clinically correct report generation from chest x-rays using templates. In *Machine learning in medical imaging: 12th international workshop, MLMI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12* (pp. 654–663). Springer.
Pino, P., Parra, D., Messina, P., Besa, C., & Uribe, S. (2020). Inspecting state of the art performance and NLP metrics in image-based medical report generation. arXiv preprint arXiv:2011.09257.
Popovici, V., Budinská, E., Dušek, L., Kozubek, M., & Bosman, F. (2017). Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. *Bioinformatics*, *33*(13), 2002–2009.
Powell, D. K., & Silberzweig, J. E. (2015). State of structured reporting in radiology, a survey. *Academic Radiology*, *22*(2), 226–233.
Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43.
Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.
Qin, H., & Song, Y. (2022). Reinforced cross-modal alignment for radiology report generation. In *Findings of the association for computational linguistics: ACL 2022* (pp. 448–458).
Radford, A. (2018). Improving language understanding by generative pre-training. Technical report.
Ramesh, V., Chi, N., & Rajpurkar, P. (2022a). Cxr-pro: Mimic-cxr with prior references omitted. *PhysioNet*. Version 1.0.0 https://doi.org/10.13026/frag-yn96
Ramesh, V., Chi, N. A., & Rajpurkar, P. (2022b). Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine learning for health* (pp. 456–473). PMLR.
Ranjit, M., Ganapathy, G., Manuel, R., & Ganu, T. (2023). Retrieval augmented chest x-ray report generation using open AI GPT models. In *Machine learning for healthcare conference* (pp. 650–666). PMLR.
Reale-Nosei, G., Amador-Domínguez, E., & Serrano, E. (2024). From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, *97*, 103264. https://doi.org/10.1016/j.media.2024.103264.
Reiner, B. I. (2009). The challenges, opportunities, and imperative of structured reporting in medical imaging. *Journal of Digital Imaging*, *22*(6), 562–568.
Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.
Ren, W., Chen, M., Qiao, Y., & Zhao, F. (2022). Global guidelines for breast cancer screening: A systematic review. *The Breast*, *64*, 85–99.
Rocha, D. M., Brasil, L. M., Lamas, J. M., Luz, G. V. S., & Bacelar, S. S. (2020). Evidence of the benefits, advantages and potentialities of the structured radiological report: An integrative review. *Artificial Intelligence in Medicine*, *102*, 101770.
Rodin, I., Fedulova, I., Shelmanov, A., & Dylov, D. V. (2019). Multitask and multimodal neural network model for interpretable analysis of x-ray images. In *2019 IEEE International conference on bioinformatics and biomedicine (BIBM)* (pp. 1601–1604). IEEE.
RSNA (2025). Radreport: A resource for structured reporting. http://www.radreport.org. Accessed: 10 January 2025.
Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., Koitka, S., Pelka, O., Abacha, A. B., G. Seco de Herrera, A. et al. (2024). Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, *11*(1), 688.
Sahni, V. A., Silveira, P. C., Sainani, N. I., & Khorasani, R. (2015). Impact of a structured report template on the quality of MRI reports for rectal cancer staging. *American Journal of Roentgenology*, *205*(3), 584–588.
Sanh, V. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
Pinto dos Santos, D., & Baeßler, B. (2018). Big data, artificial intelligence, and structured reporting. *European Radiology Experimental*, *2*(1), 42.
Pinto dos Santos, D., Brodehl, S., Baeßler, B., Arnhold, G., Dratsch, T., Chon, S.-H., Mildenberger, P., & Jungmann, F. (2019). Structured report data can be used to develop deep learning algorithms: A proof of concept in ankle radiographs. *Insights into Imaging*, *10*, 1–8.
Pinto dos Santos, D., Scheibl, S., Arnhold, G., Maehringer-Kunz, A., Düber, C., Mildenberger, P., & Kloeckner, R. (2018). A proof of concept for epidemiological research using structured reporting with pulmonary embolism as a use case. *British Journal of Radiology*, *91*(1088), 20170564. https://doi.org/10.1259/bjr.20170564.
Scalco, E., & Rizzo, G. (2017). Texture analysis of medical images for radiotherapy applications. *The British Journal of Radiology*, *90*(1070), 20160642.
Schoeppe, F., Sommer, W. H., Nörenberg, D., Verbeek, M., Bogner, C., Westphalen, C. B., Dreyling, M., Rummeny, E. J., & Fingerle, A. A. (2018). Structured reporting adds clinical value in primary CT staging of diffuse large b-cell lymphoma. *European Radiology*, *28*, 3702–3709.
Schwartz, L. H., Panicek, D. M., Berk, A. R., Li, Y., & Hricak, H. (2011). Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, *260*(1), 174–181.
Segrelles, J. D., Medina, R., Blanquer, I., & Martí-Bonmatí, L. (2017). Increasing the efficiency on producing radiology reports for breast cancer diagnosis by means of struc-

tured reports. *Methods of Information in Medicine*, *56*(03), 248–260.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).
Serra, F. D., Wang, C., Deligianni, F., Dalton, J., & O'Neil, A. Q. (2023a). Controllable chest x-ray report generation from longitudinal representations. arXiv preprint arXiv:2310.05881.
Serra, F. D., Wang, C., Deligianni, F., Dalton, J., & O'Neil, A. Q. (2023b). Finding-aware anatomical tokens for chest x-ray automated reporting. arXiv preprint arXiv:2308.15961.
Shang, C., Cui, S., Li, T., Wang, X., Li, Y., & Jiang, J. (2022). Matnet: Exploiting multi-modal features for radiology report generation. *IEEE Signal Processing Letters*, *29*, 2692–2696.
Shin, H.-C., Lu, L., Kim, L., Seff, A., Yao, J., & Summers, R. M. (2016a). Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *Journal of Machine Learning Research*, *17*(107), 1–31.
Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., & Summers, R. M. (2016b). Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2497–2506).
Sirshar, M., Paracha, M. F. K., Akram, M. U., Alghamdi, N. S., Zaidi, S. Z. Y., & Fatima, T. (2022). Attention based automated radiology report generation using CNN and LSTM. *Plos One*, *17*(1), e0262209.
Sistrom, C. L., & Honeyman-Buck, J. (2005). Free text versus structured format: Information transfer efficiency of radiology reports. *American Journal of Roentgenology*, *185*(3), 804–812.
Sloan, P., Clatworthy, P., Simpson, E., & Mirmehdi, M. (2024). Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, *18*, 368–387. https://doi.org/10.1109/RBME.2024.3408456.
Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (2020). Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167.
Sobez, L. M., Kim, S. H., Angstwurm, M., Störmann, S., Pförringer, D., Schmidutz, F., Prezzi, D., Kelly-Morland, C., Sommer, W. H., Sabel, B. et al. (2019). Creating high-quality radiology reports in foreign languages through multilingual structured reporting. *European Radiology*, *29*, 6038–6048.
Song, X., Zhang, X., Ji, J., Liu, Y., & Wei, P. (2022). Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th international conference on computational linguistics* (pp. 2388–2397).
Song, Y., Hua, X., Zhang, K., Zan, H., & Li, R. (2024). Multi-granularity semantic guided transformer for radiology report generation. In *CCF international conference on natural language processing and Chinese computing* (pp. 458–471). Springer.
Sudirman, S., Al Kafri, A., Natalia, F., Meidia, H., Afriliana, N., Al-Rashdan, W., Bashtawi, M., & Al-Jumaily, M. (2019a). Lumbar spine MRI dataset. Mendeley Data, V2, https://doi.org/10.17632/k57fr854j2.2.
Sudirman, S., Al Kafri, A., Natalia, F., Meidia, H., Afriliana, N., Al-Rashdan, W., Bashtawi, M., & Al-Jumaily, M. (2019b). Radiologists notes for lumbar spine MRI dataset. Mendeley Data, V2, https://doi.org/10.17632/s6bgczr8s2.2.
Sun, L., Wang, W., Li, J., & Lin, J. (2019). Study on medical image report generation based on improved encoding-decoding method. In *Intelligent computing theories and application: 15th international conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I 15* (pp. 686–696). Springer.
Sun, S., Mei, Z., Li, X., Tang, T., Su, Z., & Wu, Y. (2024). A label information fused medical image report generation framework. *Artificial Intelligence in Medicine*, *150*, 102823.
Syeda-Mahmood, T., Wong, K. C. L., Gur, Y., Wu, J. T., Jadhav, A., Kashyap, S., Karargyris, A., Pillai, A., Sharma, A., Syed, A. B. et al. (2020). Chest x-ray report generation through fine-grained label learning. In *Medical image computing and computer assisted intervention–MICCAI 2020: 23rd international conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23* (pp. 561–571). Springer.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
Tang, Y., Yang, H., Zhang, L., & Yuan, Y. (2024). Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, *237*, 121442.
Tanwani, A. K., Barral, J., & Freedman, D. (2022). Repsnet: Combining vision with language for automated medical reports. In *International conference on medical image computing and computer-assisted intervention* (pp. 714–724). Springer.
Tempero, M. A., Malafa, M. P., Al-Hawary, M., Behrman, S. W., Benson, A. B., Cardin, D. B., Chiorean, E. G., Chung, V., Czito, B., Del Chiaro, M. et al. (2021). Pancreatic adenocarcinoma, version 2.2021, NCCN clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, *19*(4), 439–457.
The Royal College of Radiologists (2023). Clinical radiology workforce census 2023. Accessed: 07 January 2025 https://www.rcr.ac.uk/media/5befglss/rcr-census-clinical-radiology-workforce-census-2023.pdf.
Tian, J., Li, C., Shi, Z., & Xu, F. (2018). A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11* (pp. 702–710). Springer.
Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56.
Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.
Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).
Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).
Wang, F., Du, S., & Yu, L. (2024a). Hergen: Elevating radiology report generation with longitudinal data. In *European conference on computer vision* (pp. 183–200). Springer.
Wang, F., Liang, X., Xu, L., & Lin, L. (2020). Unifying relational sentence generation and retrieval for medical image report composition. *IEEE Transactions on Cybernetics*, *52*(6), 5015–5025.
Wang, H., Niu, J., Liu, X., & Wang, Y. (2023a). A doctors behavior aware and domain knowledge driven model for medical report generation. In *2023 IEEE International conference on bioinformatics and biomedicine (BIBM)* (pp. 2687–2694). IEEE.
Wang, X., Figueredo, G., Li, R., Zhang, W. E., Chen, W., & Chen, X. (2024b). A survey of deep learning-based radiology report generation using multimodal data. arXiv preprint arXiv:2405.12833.
Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).
Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9049–9058).
Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Li, C., & Tang, J. (2024c). Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. arXiv preprint arXiv:2410.00379.
Wang, Y., Lin, Z., Xu, Z., Dong, H., Luo, J., Tian, J., Shi, Z., Huang, L., Zhang, Y., Fan, J. et al. (2024d). Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, *578*, 127374.
Wang, Z., Han, H., Wang, L., Li, X., & Zhou, L. (2022a). Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, *41*(10), 2803–2813.
Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023b). Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11558–11567).
Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023c). R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, *1*(3), 100033.
Wang, Z., Tang, M., Wang, L., Li, X., & Zhou, L. (2022b). A medical semantic-assisted transformer for radiographic report generation. In *International conference on medical image computing and computer-assisted intervention* (pp. 655–664). Springer.
Wang, Z., Zhou, L., Wang, L., & Li, X. (2021). A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2433–2442).
Weiss, D. L., & Bolos, P. R. (2010). Reporting and dictation. *Practical imaging informatics: Foundations and applications for PACS professionals*, (pp. 147–162).
Wu, C., Lei, J., Zheng, Q., Zhao, W., Lin, W., Zhang, X., Zhou, X., Zhao, Z., Zhang, Y., Wang, Y. et al. (2023a). Can GPT-4V (ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. arXiv preprint arXiv:2310.09909.
Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023b). Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463.
Wu, J. T., Agu, N. N., Lourentzou, I., Sharma, A., Paguio, J. A., Yao, J. S., Dee, E. C., Mitchell, W., Kashyap, S., Giovannini, A. et al. (2021). Chest imagenome dataset (version 1.0. 0). *PhysioNet*, *5*(18), 2–3.
Wu, P., Dong, H., Lin, Y., Ding, Y., & Peng, Y. (2026). A disease-aware dual-stage framework for chest x-ray report generation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 33953–33961). (*vol. 40*).
Xie, X., Xiong, Y., Yu, P. S., Li, K., Zhang, S., & Zhu, Y. (2019). Attention-based abnormal-aware fusion network for radiology report generation. In *Database systems for advanced applications: DASFAA 2019 International workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22–25, 2019, Proceedings 24* (pp. 448–452). Springer.
Xiong, Y., Du, B., & Yan, P. (2019). Reinforced transformer for medical image captioning. In *Machine learning in medical imaging: 10th international workshop, MLMI 2019, held in conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10* (pp. 673–680). Springer.
Xu, K. (2015). Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.
Xu, Z., Xu, W., Wang, R., Chen, J., Qi, C., & Lukasiewicz, T. (2023). Hybrid reinforced medical report generation with m-linear attention and repetition penalty. *IEEE Transactions on Neural Networks and Learning Systems* , *36*(2), 2206–2220. https://doi.org/10.1109/TNNLS.2023.3343391.
Xue, Y., & Huang, X. (2019). Improved disease classification in chest x-rays with transferred features from report generation. In *Information processing in medical imaging: 26th International conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26* (pp. 125–138). Springer.
Xue, Y., Tan, Y., Tan, L., Qin, J., & Xiang, X. (2024). Generating radiology reports via auxiliary signal guidance and a memory-driven network. *Expert Systems with Applications*, *237*, 121260.
Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In *Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I* (pp. 457–466). Springer.
Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., & Hsu, C.-N. (2021). Weakly supervised contrastive learning for chest x-ray report generation. arXiv

preprint arXiv:2109.12242.
Yan, B., Pei, M., Zhao, M., Shan, C., & Tian, Z. (2022). Prior guided transformer for accurate radiology reports generation. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5631–5640.
Yan, K., Wang, X., Lu, L., & Summers, R. M. (2018). Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, *5*(3), 036501–036501.
Yan, S., Cheung, W. K., Chiu, K., Tong, T. M., Cheung, K. C., & See, S. (2023). Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, *42*(8), 2211–2222.
Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–19.
Yang, S., Ji, J., Zhang, X., Liu, Y., & Wang, Z. (2021a). Weakly guided hierarchical encoder-decoder network for brain ct report generation. In *2021 IEEE International conference on bioinformatics and biomedicine (BIBM)* (pp. 568–573). IEEE.
Yang, S., Niu, J., Wu, J., Wang, Y., Liu, X., & Li, Q. (2021b). Automatic ultrasound image report generation with adaptive multimodal attention mechanism. *Neurocomputing*, *427*, 40–49.
Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., & Xiao, L. (2023a). Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, *86*, 102798.
Yang, S., Wu, X., Ge, S., Zhou, S. K., & Xiao, L. (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, *80*, 102510.
Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., & Xie, P. (2020). Covid-CT-dataset: A CT scan dataset about covid-19. arXiv preprint arXiv:2003.13865.
Yang, X., Ye, M., You, Q., & Ma, F. (2021c). Writing by memorizing: Hierarchical retrieval-based medical report generation. arXiv preprint arXiv:2106.06471.
Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023b). The dawn of lmms: Preliminary explorations with GPT-4V (ision), *9*(1), 1. arXiv preprint arXiv:2309.17421
Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision* (pp. 4894–4902).
Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., & Zheng, Q. (2019). Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE International conference on data mining (ICDM)* (pp. 728–737). IEEE.
Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series* (pp. 022022). IOP Publishing (*Vol. 1168*).
You, D., Liu, F., Ge, S., Xie, X., Zhang, J., & Wu, X. (2021). Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24* (pp. 72–82). Springer.
You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651–4659).
Yousem, D. M. (2019). In opposition to standardized templated reporting. *Academic Radiology*, *26*(7), 981–982.
Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Lee, H. M. H., Abad, Z. S. H., Ng, A. Y. et al. (2022). Evaluating progress in automatic chest x-ray radiology report generation.. *medrxiv*, *31*, 10–1016. Preprint posted online August,
Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(12), 4467–4480.
Yuan, J., Liao, H., Luo, R., & Luo, J. (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22* (pp. 721–729). Springer.
Zeng, X., Liao, T., Xu, L., & Wang, Z. (2024). Aermnet: Attention-enhanced relational memory network for medical image report generation. *Computer Methods and Programs in Biomedicine*, *244*, 107979.
Zeng, X., Wen, L., Liu, B., & Qi, X. (2020a). Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*, *392*, 132–141.
Zeng, X., Wen, L., Xu, Y., & Ji, C. (2020b). Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. *Computer Methods and Programs in Biomedicine*, *197*, 105700.
Zhang, D., Ren, A., Liang, J., Liu, Q., Wang, H., & Ma, Y. (2022). Improving medical x-ray report generation by using knowledge graph. *Applied Sciences*, *12*(21), 11111.
Zhang, J., Cheng, M., Cheng, Q., Shen, X., Wan, Y., Zhu, J., & Liu, M. (2024a). Hierarchical medical image report adversarial generation with hybrid discriminator. *Artificial Intelligence in Medicine*, *151*, 102846.
Zhang, J., Shen, X., Wan, S., Goudos, S. K., Wu, J., Cheng, M., & Zhang, W. (2023a). A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE Journal of Biomedical and Health Informatics*, *27*(10), 5110–5121.
Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019a). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.
Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., & Xie, W. (2024b). Radgenome-chest CT: A grounded vision-language dataset for chest CT analysis. arXiv preprint arXiv:2404.16754.
Zhang, Y., Merck, D., Tsai, E. B., Manning, C. D., & Langlotz, C. P. (2019b). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. arXiv preprint arXiv:1911.02541.
Zhang, Y., Wang, X., Guo, Z., & Li, J. (2018). Imagesem at imageCLEF 2018 caption task: Image retrieval and transfer learning. CLEF Working Notes. ImageCLEF 2018 caption task paper.
Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., & Xu, D. (2020). When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12910–12917). (*Vol. 34*).
Zhang, Y., Yuan, Y., & Yao, A. C.-C. (2023b). Meta prompting for AI systems. arXiv preprint arXiv:2311.11482.
Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., & Langlotz, C. P. (2021). Biomedical and clinical english model packages for the stanza python NLP library. *Journal of the American Medical Informatics Association*, *28*(9), 1892–1899.
Zhang, Z., Xie, Y., Xing, F., McGough, M., & Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6428–6436).
Zhao, G., Yan, Y., & Zhao, Z. (2023a). Normal-abnormal decoupling memory for medical report generation. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 1962–1977).
Zhao, G., Zhao, Z., Gong, W., & Li, F. (2023b). Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. *Artificial Intelligence in Medicine*, *146*, 102714.
Zhao, K., Xiao, C., Yan, S., Tang, H., Cheung, W. K., Moubayed, N. A., Zhan, L., & Lin, C. (2024). X-ray made simple: Lay radiology report generation and robust evaluation. arXiv preprint arXiv:2406.17911.
Zheng, F., Li, M., Wang, Y., Yu, W., Wang, R., Chen, Z., Xiao, N., & Lu, Y. (2024). Intensive vision-guided network for radiology report generation. *Physics in Medicine & Biology*, *69*(4), 045008.
Zhou, H.-Y., Chen, X., Zhang, Y., Luo, R., Wang, L., & Yu, Y. (2022). Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, *4*(1), 32–40.