mathematics

MDPI

Article

MedChain: A Knowledge-Guided Evidence Reasoning Framework for Automated Radiology Report Generation

Kangda Cheng [1], Weilin Wu [1,*], Hebin Wang [2], Ziyao Ding [2] and Jinlong Liu [2]

[1] College of Computer Engineering, Jimei University, Xiamen 361021, China; chengkangda@jmu.edu.cn
[2] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; 25s005057@stu.hit.edu.cn (H.W.); 25b905018@stu.hit.edu.cn (Z.D.); yq20@hit.edu.cn (J.L.)
* Correspondence: weilinwu@jmu.edu.cn

Abstract

Automated radiology report generation aims to produce clinically coherent reports from medical images. However, existing methods typically formulate this task as a direct mapping from visual features to textual descriptions, overlooking the medical observation, evidence identification, and clinical reasoning processes involved in radiologists' report generation. Due to the lack of explicit medical evidence representations and knowledge-enhanced semantic association modeling mechanisms, existing models often struggle to establish reliable associations between imaging abnormalities and disease concepts, resulting in reports that may lack sufficient imaging support or medical consistency. To address these limitations, we propose MedChain, a medical report generation framework based on a structured reasoning pathway. Unlike conventional methods, we formulate medical report generation as a multi-stage medical reasoning process. Specifically, we first construct normal reference representations using normal training cases and extract difference-aware clinical evidence by modeling the representation discrepancy between patient visual features and normal reference patterns. Subsequently, we employ a medical reasoning graph to encode structured relationships among anatomical structures, imaging findings, and disease concepts, and fuse the knowledge-enhanced medical concept representations with visual evidence through cross-modal interaction. Furthermore, we introduce a hierarchical clinical report generation strategy that simulates the real radiological reporting workflow by decomposing report generation into two stages: findings generation and impression generation. This strategy enhances the structural consistency and clinical logic of generated reports. Experiments on IU X-Ray and MIMIC-CXR evaluate report generation performance, clinical-label agreement, and cross-domain transfer.

Keywords: medical image report generation; vision–language learning; radiology report generation; multimodal learning

MSC: 68T07; 92C55

Check for updates

Academic Editor: Romualdas Bausys

Received: 27 September 2026
Revised: 6 November 2026
Accepted: 19 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Chest radiography is widely used to evaluate the lungs, pleural spaces, and cardiomediastinal structures for potential abnormalities. Interpreting chest radiographs entails identifying imaging findings, localizing them anatomically, and distinguishing observed signs from diagnostic impressions. The volume of chest X-ray examinations places substantial demands on radiologists for image interpretation and report documentation [1,2].

Automated chest X-ray report generation therefore seeks to convert such observations into coherent, clinically meaningful reports [3–5].

Deep learning has advanced automated chest X-ray report generation. Early image-captioning methods adopted an encoder–decoder framework. They extracted image features with convolutional neural networks (CNNs) and generated descriptions using recurrent neural networks (RNNs) [6]. This image-to-text paradigm subsequently provided a foundation for chest radiograph reporting. Later chest radiograph reporting methods incorporated attention mechanisms [7], knowledge enhancement [8], and joint vision–language modeling [9]. These mechanisms help associate radiographic features with clinically relevant concepts. R2Gen [10] introduces a relational memory into the decoder to capture recurring report patterns, whereas Park et al. [11] use feature differences between patient and normal images together with tag information. Neither method learns a bank of normal visual prototypes to retrieve token-specific references for difference-aware evidence extraction. Recently, multimodal large language models and vision–language models such as CXR-LLaVA [2] have extended chest X-ray interpretation to free-text report generation, although missed abnormalities and unsupported diagnostic statements remain concerns. However, generating reports from chest radiographs requires more than visual description. It is a complex process involving image observation, medical concept association, and structured information organization. In real clinical settings, radiologists do not directly form final reports from images. Instead, they follow a step-by-step observation and analysis process. First, physicians observe anatomical structures and identify potential abnormalities. They then formulate specific imaging findings based on these abnormal regions. Next, they use medical knowledge to analyze the relationships between imaging manifestations and potential diseases. Finally, they organize the diagnostic information into a standard radiology report structure [12].

Current methods for chest X-ray reporting often skip these intermediate reasoning steps. They directly generate text using visual features. This design leads to two key limitations. First, existing methods lack explicit medical evidence representations. The high-dimensional features from traditional visual encoders contain abundant imaging information. However, they remain fundamentally abstract visual representations. Without a medical evidence layer, models tend to rely on statistical language patterns in the training data. They fail to reason based on actual imaging evidence. Second, current methods lack a medical reasoning process connecting visual manifestations to disease concepts. Chest radiographic findings can be nonspecific. For example, a “lung opacity” might correspond to multiple conditions, such as pneumonia, atelectasis, pulmonary edema, or even a tumor. A reliable model must understand the complex medical associations between abnormal findings and diseases. It should not merely learn simple correlations between visual patterns and disease labels.

To address these issues, we propose MedChain. It is a novel chest X-ray report generation framework based on a structured reasoning pathway. Unlike existing methods, MedChain formulates report generation as a multi-stage process guided by a clinical reasoning chain. Specifically, MedChain first introduces a clinical evidence reasoning module (CERM), which models potential abnormality-related visual evidence by comparing patient image representations with normal anatomical reference representations learned from normal cases. By using representations learned from normal chest radiographs as references, this module enables the model to focus more on regional deviations from normal visual patterns rather than relying solely on global visual representations of patient images. Subsequently, we propose a medical reasoning-aware report generator, which incorporates structured relationships among anatomical structures, radiological findings, and disease concepts through a medical knowledge graph. The knowledge-enhanced medical concept

representations then interact with visual evidence to provide medical semantic information for report generation. Finally, we adopt a hierarchical clinical report generation strategy that decomposes the report generation process into two stages: findings generation and impression generation. This design enables the generation process to better align with the actual workflow of radiological report writing.

The main contributions of this work are summarized as follows:

(1) We propose a normality-aware clinical evidence reasoning module for chest X-ray report generation. By constructing a normal-reference memory from normal training cases, the proposed module models deviations between patient visual representations and normal reference patterns, providing difference-aware visual evidence for subsequent medical semantic modeling and report generation.

(2) We develop a knowledge-guided medical reasoning module that incorporates structured relations among anatomical structures, radiological observations, and disease concepts. The resulting knowledge-enhanced concept representations interact with visual evidence to provide an intermediate medical semantic space for chest radiograph reporting.

(3) We introduce a hierarchical clinical report generation strategy that decomposes chest X-ray report generation into findings generation followed by impression generation. By conditioning the diagnostic impression on the generated findings and medical semantic representations, the proposed strategy explicitly models the dependency between imaging observations and diagnostic summaries.

2. Related Work

2.1. Vision–Language-Based Medical Report Generation

Recently, inspired by natural image captioning, encoder–decoder architectures have been widely applied to medical report generation. Early representative works combined CNNs and RNNs for image-to-text mapping [13]. Other methods introduced multi-feature fusion and attention mechanisms to enhance visual representations [14,15]. With the development of Transformer architectures, pure Transformer-based models and memory-augmented models have been proposed to capture long-range dependencies in radiology reports and model global context [16,17].

Researchers have further exploited the strong long-range dependency modeling capability of Transformers to improve medical report generation. Align Transformer [9] learns disease-related visual representations through hierarchical alignment between visual regions and disease labels, thereby alleviating the problem of sparse abnormal regions. AIMNet [18] further introduces an adaptive image–label fusion strategy that dynamically integrates visual information with disease-label information to improve the description of abnormal regions. METransformer [19] introduces multiple learnable expert tokens, allowing different experts to attend to different image regions. Information exchange among these experts further enhances the visual representations. Liu et al. [20] investigated multi-scale and boundary-aware representations for robust human parsing in complex scenes.

Although these methods have substantially improved medical report generation performance, they still rely primarily on implicit associations between visual feature learning and text generation. Medical report generation requires not only describing image content but also inferring potential disease states from radiological findings. Therefore, a direct vision–language mapping is insufficient to fully model the clinical diagnostic process. In addition, existing methods typically use raw image features or disease labels as inputs but do not explicitly model how patient images deviate from normal anatomical states. As a result, it remains difficult for these models to distinguish clinically meaningful abnormal evidence from irrelevant visual patterns.

2.2. Knowledge-Enhanced and Reasoning-Driven Medical Report Generation

To overcome the limitations of direct vision-to-text mapping, recent studies have explored knowledge-enhanced and structured reasoning approaches to establish more explicit connections among visual information, medical concepts, and diagnostic conclusions. A representative research direction is the incorporation of external medical knowledge into report generation. Zhang et al. [21] first investigated the role of knowledge graphs in medical report generation. They constructed a disease–finding relation graph and employed graph convolutional networks to learn structural relationships among disease entities, thereby introducing medical knowledge into the report generation process. This study highlighted that, unlike natural image captioning, medical report generation places greater emphasis on the relationships between disease-related keywords and their attributes, making medical knowledge essential for improving model understanding. KiUT [22] further proposed a knowledge-injected U-Transformer framework, which integrates clinical knowledge into the vision–language interaction process by constructing a symptom graph and a knowledge distiller, thereby improving the medical consistency of generated reports.

Another line of research focuses on cross-modal semantic alignment between visual information and textual descriptions. The Reinforced Cross-modal Alignment method [23] pointed out that conventional attention mechanisms struggle to learn accurate vision–language correspondences due to the lack of region-level alignment annotations between medical images and reports. Therefore, this method introduces a cross-modal memory to capture latent associations between visual and textual modalities and employs reinforcement learning to optimize the cross-modal alignment process. Although this approach enhances information interaction between images and texts [24], its primary objective remains improving modality matching rather than establishing a medical reasoning pathway from imaging evidence to disease diagnosis.

Recently, researchers have focused on the structural characteristics of medical reports. They attempt to simulate the real clinical reporting workflow. Hou et al. [25] proposed an observation-guided report generation framework that generates an observation plan and incorporates an observation graph with a tree reasoning mechanism to enrich medical observation information, making the report generation process closer to clinical reasoning.

Furthermore, with the emergence of large vision–language models, recent studies have investigated the use of large-scale pre-trained models to improve medical image understanding [26]. CXR-LLaVA [2] trains a visual encoder on large-scale chest X-ray datasets and combines it with a large language model for chest X-ray understanding and report generation, demonstrating the potential of multimodal large models in medical vision–language tasks. However, large-model-based approaches still rely on the implicit associations between visual representations and language generation, and may suffer from medical hallucinations, insufficient abnormality localization, and limited interpretability of reasoning processes in complex cases.

Overall, existing medical report generation methods have gradually evolved from early vision–language mapping approaches toward knowledge enhancement, cross-modal alignment, and structured reasoning. However, most existing studies optimize only individual aspects, such as visual representation learning, medical knowledge integration, or report structure modeling, while lacking a unified framework that simultaneously connects abnormal imaging evidence, medical knowledge relationships, and the clinical report generation process. Moreover, existing methods generally model patient image features directly while overlooking the variations between patient images and normal anatomical states, limiting their ability to capture clinically meaningful abnormal evidence.

To address these limitations, this paper proposes a medical image report generation framework based on a structured clinical reasoning pathway. Unlike existing approaches,

our method first employs normality-aware clinical evidence reasoning to model the differences between patient images and normal states, thereby obtaining more clinically meaningful abnormal evidence. It then utilizes a medical reasoning graph to establish structured relationships among visual findings, medical concepts, and disease states. Finally, a hierarchical report generation strategy is adopted to simulate the complete reasoning process of radiologists, from image observation to diagnostic summarization.

3. Methodology

3.1. Overview

An overview of the proposed structured reasoning pathway-based medical image report generation framework (MedChain) is presented in Figure 1. Given an input image, the proposed framework first extracts clinically meaningful visual evidence through the clinical evidence reasoning module. Unlike conventional visual encoders that primarily capture visual content, this module performs normality-aware contrastive reasoning to identify deviations between patient images and normal anatomical reference states, thereby highlighting potential pathological regions. The extracted evidence is then fed into the medical reasoning-aware report generator, which incorporates medical knowledge to establish associations between visual evidence and disease concepts. Following the clinical report generation workflow, the generator sequentially produces imaging findings and diagnostic impressions, ultimately generating a complete medical report.

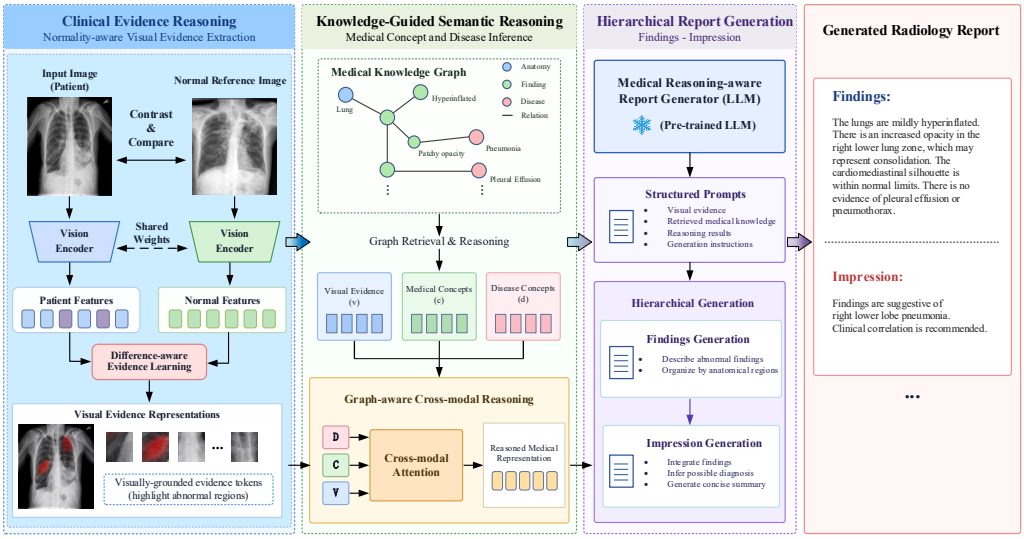

Figure 1. Overview of the proposed structured reasoning pathway-based medical image report generation framework (MedChain). Given a chest radiograph, MedChain first extracts difference-aware visual evidence by comparing patient visual representations with normal reference representations. Then, a medical reasoning graph models structured relationships among anatomical structures, radiological findings, and disease concepts, producing knowledge-enhanced medical representations. These representations interact with visual evidence through cross-modal attention to obtain patient-specific semantic representations. Finally, the hierarchical report generation module follows the clinical reporting workflow by generating findings first and subsequently producing impression conditioned on the generated findings and medical semantic representations.

3.2. Clinical Evidence Reasoning Module

Given an input image I, a hierarchical visual encoder is first used to extract multi-scale visual representations:

$$F_I = \{F_I^1, F_I^2, \ldots, F_I^L\} \quad (1)$$

where L denotes the number of stages in the visual encoder, and F_I^1 represents the feature output from the first stage. Features from different stages contain different levels of semantic information. Low-level features capture local textures and edge patterns, which help identify small-scale abnormalities. In contrast, deeper features contain more abstract anatomical structures and semantic information, facilitating the understanding of disease-related regions. Therefore, we retain multi-scale visual representations and project features from different levels into a unified feature space through linear mapping:

$$Z_l = W_l F_I^l \tag{2}$$

where W_l denotes the feature projection matrix of the l-th layer. Subsequently, all multi-scale features are flattened into visual tokens:

$$V_I = Concat(Z_1, Z_2, \ldots, Z_L) \tag{3}$$

where

$$V_I = \{v_1, v_2, \ldots, v_N\} \tag{4}$$

Each token captures contextual visual information from a local image region. However, patient image features alone cannot explicitly measure the degree of deviation between local visual patterns and normal anatomical states. Therefore, we further construct a normal-reference representation space, where visual prototypes learned from normal cases in the training set are used to provide normal-state visual references for patient image features.

We assume that normal chest X-rays exhibit relatively stable patterns in the visual representation space. Thus, we learn a set of normal visual prototypes from report-filtered candidate normal cases. These prototypes approximate the normal visual feature space. We strictly use only training set data to construct this normal reference memory bank. This prevents any information leakage from the validation or test sets.

Specifically, we initialize and learn a normal reference memory bank $M_N = \{m_1, m_2, \ldots, m_K\}$, where m_k denotes the k-th learnable normal prototype. These prototypes represent typical visual patterns of normal cases within a shared feature space. They are jointly optimized with the entire model during training. Each normal reference is an attention-weighted combination of learnable prototypes rather than a stored normal radiograph.

The normal-reference prototypes are initialized from the candidate normal examinations in the training partition. The ImageNet-pretrained Swin encoder and Xavier-uniform-initialized projection layers produce 512-dimensional multiscale visual tokens for these examinations. Euclidean K-means with K-means++ seeding is then fitted to the training-normal tokens using 64 clusters, 10 initializations, a maximum of 300 iterations, a convergence tolerance of 0.0001, and random seed 42. The resulting 64 cluster centroids initialize the rows of the normal-reference memory matrix. These prototype vectors remain learnable and are subsequently updated by backpropagation jointly with the report-generation model. Neither validation nor test features participate in prototype initialization.

For an input visual token v_i, the model finds its corresponding normal state using an attention mechanism:

$$\alpha_{ik} = \frac{exp(v_i^T m_k)}{\sum_{j=1}^{K} exp(v_i^T m_j)} \tag{5}$$

where α_{ik} represents the correlation between the current region and the k-th normal prototype.

Accordingly, the normal reference representation corresponding to this region is obtained as

$$n_i = \sum_{k=1}^{K} \alpha_{ik} m_k \tag{6}$$

These regional representations form the normal reference sequence $V_N = n_1, n_2, ..., n_N$. For the patient visual representation V_I, we then compute its residual representation as

$$D = V_I - V_N \tag{7}$$

where D characterizes the representation deviation between patient visual features and normal reference patterns. It should be noted that this residual representation does not directly indicate pathological abnormalities. Instead, it serves as an auxiliary signal for the subsequent evidence attention module to identify potentially abnormal information. However, directly using difference features still has limitations. Since different regions contribute unequally to clinical diagnosis, simple subtraction cannot determine which changes are more diagnostically significant. For example, minor texture variations in the lung and obvious effusion regions might yield similar deviation magnitudes. Therefore, we further introduce a difference-aware attention mechanism, enabling the model to adaptively select regions with greater diagnostic relevance.

Specifically, we first generate the query vector $Q = W_q V_I$ from the patient visual representation. Next, we generate a key vector $K = W_k D$ from the difference representation. The value vector $V = W_v V_I$ is obtained from the patient visual features. The regional correlation is then calculated as

$$A = softmax(\frac{QK^T}{\sqrt{d}}) \tag{8}$$

where d denotes the feature dimension. Finally, the enhanced medical evidence representation is obtained as

$$E = AV \tag{9}$$

where $E = \{e_1, e_2, ..., e_N\}$ represents the medical visual evidence learned by the model. Unlike conventional visual features, E contains not only imaging information but also the variation in patient states relative to normal anatomical states. Therefore, this module reweights visual information based on the representation discrepancy between patient images and normal references, enabling subsequent medical reasoning and report generation to focus more on potentially abnormal visual evidence.

Normal-reference prototypes learned from report-filtered examinations may incorporate patterns associated with unreported mild abnormalities. Because the normal reference is retrieved through similarity-weighted attention over learnable prototypes, such contamination can alter both the prototype representations and their retrieval weights, thereby changing the reference used for difference-aware evidence extraction.

To illustrate this feature-level effect, consider a fixed retrieved reference $\tilde{n}_i = (1 - \rho)u_i + \rho a_i$, where u_i and a_i denote distinct normal and abnormal feature patterns and ρ is the effective contribution of the abnormal pattern. For a patient feature matching a_i, the residual retains $1 - \rho$ of its uncontaminated magnitude. Conversely, a normal feature matching u_i acquires a spurious residual with a normalized magnitude of ρ. Both effects can weaken lesion-related evidence or emphasize non-pathological regions in subsequent difference-aware attention.

The mixture weight represents abnormal-pattern content in a fixed retrieved reference rather than the proportion of mislabeled examinations. These identities describe feature-level failure mechanisms under a fixed-reference mixture assumption, whereas MedChain uses query-dependent retrieval and jointly optimized prototypes.

3.3. Medical Reasoning-Aware Report Generator

Unlike conventional visual classification tasks, the relationship between visual manifestations and diseases in medical images is rarely one-to-one. Instead, complex many-to-many

associations often exist between imaging findings and disease concepts. Therefore, medical knowledge structures are required to explicitly describe these relationships. In this work, we construct a medical reasoning graph $G = (V, E_g)$, where V denotes the set of medical concept nodes and E_g represents the set of medical relations. The nodes mainly consist of three types of medical entities: anatomical structures, radiological findings, and disease concepts. The edges represent semantic relationships among these entities.

Assume that the medical graph contains M nodes. We denote their initial node representations as $H^0 = h_1^0, h_2^0, \ldots, h_M^0$, where h_i^0 denotes the initial state of the i-th medical concept node. It is initialized using a learnable embedding corresponding to that concept.

To construct the medical knowledge structure, we first extract structured medical relations from radiology reports. As illustrated in Figure 2, an LLM-based medical concept extractor converts free-form radiology descriptions into structured triples, including modification, anatomical localization, and disease association relations. These triples are subsequently organized into a medical reasoning graph, where nodes represent clinical concepts and edges represent semantic relationships.

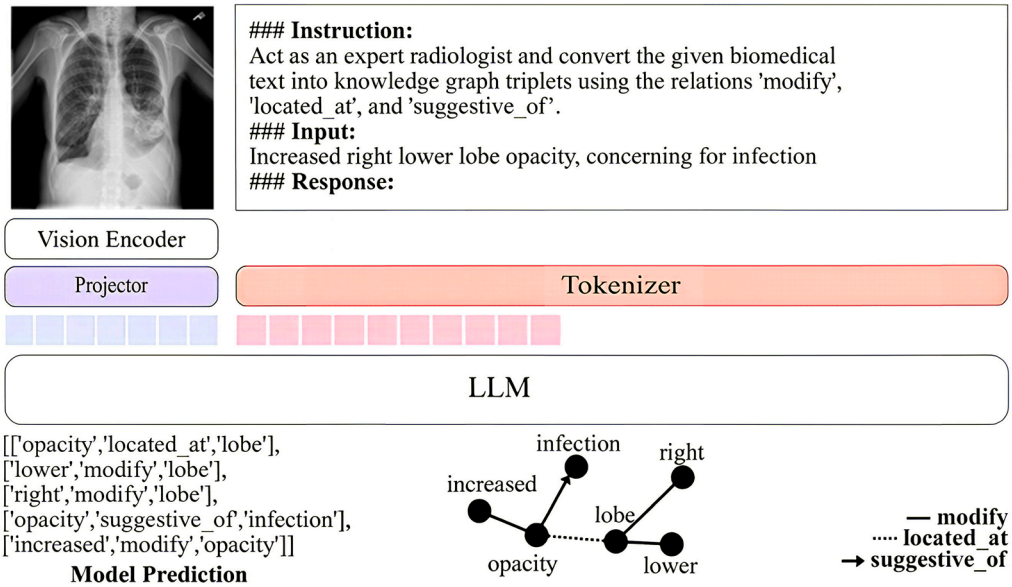

Figure 2. Construction of the medical reasoning graph from radiology text. An LLM-based extractor converts radiological descriptions into structured medical triples, including modification, anatomical location, and disease suggestion relations. The extracted triples are further organized into a graph representation to provide medical knowledge for evidence-guided report generation.

Medical triples are extracted from training-set radiology reports using the post-trained Qwen3.5-9B checkpoint Qwen/Qwen3.5-9B. This is a dense, 9-billion-parameter vision–language model with a 32-layer language backbone, organized into eight repetitions of three Gated DeltaNet layers followed by one gated-attention layer [27]. The image pathway in Figure 2 is illustrative; graph construction uses report text only. The checkpoint is kept fixed, thinking output is disabled, and extraction uses zero-shot greedy decoding with a maximum of 512 output tokens. Each training report is processed separately with the following fixed prompt template.

Instruction: Act as an expert radiologist and extract directed knowledge-graph triples from the Input radiology report. Return a list of [source, relation, target] triples using only modify, located_at, and suggestive_of. For modify, the source is an attribute and the target is the modified concept. For located_at, the source is a radiographic finding and the target is its anatomical location. For suggestive_of, the source is a finding and the target

is a possible disease. Preserve laterality and anatomical qualifiers. Omit negated findings and unsupported relations. Represent explicitly uncertain disease associations only through suggestive_of, without asserting them as confirmed diagnoses. Use lowercase entity strings. Return the list only, or [] if no valid relation is present.

Input: {report_text}
Response:

The returned triples are parsed as ordered source-relation-target lists and mapped into a canonical medical concept vocabulary before graph assembly. Normalization applies Unicode NFKC conversion, lowercasing, removal of redundant punctuation and whitespace, singularization of radiological finding terms, and deterministic abbreviation and synonym mapping against the training-corpus concept dictionary. For example, opacities is mapped to opacity and RLL to right lower lobe. Anatomical laterality and lobar qualifiers are retained, while descriptive attributes such as increased and lower can be represented as separate modifier nodes. Triples with unsupported relation labels, unmapped entities, or invalid edge directions are rejected. Negated observations are excluded from positive graph edges; explicitly suggested diagnoses retain the suggestive_of relation without being treated as confirmed findings. Exact duplicate directed triples are removed, and the resulting canonical entities and edges form the graph used for medical concept propagation. Only training-set reports contribute to this graph.

Each extracted medical triple becomes a directed edge in the reasoning graph and determines which concept nodes exchange information during graph attention. A falsely extracted anatomical localization changes the anatomical context of an observation, whereas an incorrect disease association can introduce a potentially unsupported clinical link. An omitted modification relation removes information about an observation or anatomical attribute. These errors alter the graph neighborhood and the messages available to subsequent Transformer layers, so downstream cross-modal attention cannot be assumed to eliminate them.

During the medical reasoning process, a graph Transformer is employed to propagate information over the medical relation graph:

$$H^{l+1} = Transformer_G(H^l, G) \tag{10}$$

where H^l denotes the representations of medical concept nodes at the l-th layer. Through graph-based information propagation, each medical concept node aggregates relevant information from its neighboring nodes, resulting in knowledge-enhanced medical concept representations:

$$C = H^{L_g} = \{c_1, c_2, \ldots, c_M\} \tag{11}$$

where L_g denotes the number of graph Transformer layers, and c_i represents the i-th medical concept representation after medical relation propagation. Compared with directly generating text from visual features, the knowledge-enhanced medical concept representations provide an intermediate semantic space that better conforms to clinical reasoning logic. This enables the model to leverage structured relationships among medical entities to facilitate report generation.

To further establish the associations between visual evidence and medical concepts, we employ a cross-modal attention mechanism to fuse the medical visual evidence representation E with the knowledge-enhanced medical concept representation C:

$$C_E = CrossAttn(Q = E, K = C, V = C) \tag{12}$$

where C_E denotes the medical semantic representation adaptively selected according to the current patient-specific imaging evidence. This process does not rely on predefined corre-

spondences between visual regions and medical concepts. Instead, the attention mechanism learns latent associations between visual evidence and relevant medical concepts.

Cross-modal attention assigns learned compatibility weights to visual evidence and graph concepts, but it does not enforce the coverage of every distinct abnormal region or require a separate visual observation for each disease hypothesis. In examinations with multiple abnormalities, highly weighted evidence can have greater influence during decoding, while related disease concepts can enter the semantic representation even when their patient-specific visual support is weak.

Subsequently, we design a hierarchical clinical report generation strategy. Medical imaging reports typically follow a predefined organizational structure rather than being simple natural language descriptions. For example, in chest X-ray reports, radiologists usually describe imaging findings first and then provide a diagnostic impression. Therefore, instead of directly generating the entire report as a single sequence, we simulate the clinical report-writing process and decompose generation into two stages. First, the model generates imaging findings based on visual evidence:

$$Y_f = \left\{ y_1^f, y_2^f, \ldots, y_m^f \right\} \tag{13}$$

with the conditional probability defined as $P(Y_f|E)$. Subsequently, the model generates the diagnostic impression by incorporating the generated findings and the medical semantic representation:

$$Y_i = \left\{ y_1^i, y_2^i, \ldots, y_n^i \right\} \tag{14}$$

with the generation probability formulated as $P(Y_i|Y_f, C_E)$. Here, C_E represents the knowledge-enhanced medical semantic representation fused with visual evidence. The final medical report is obtained by concatenating the outputs from the two stages according to the standard report structure:

$$Y = \left[Y_f; Y_i \right] \tag{15}$$

where $[;]$ denotes the sequential concatenation of the findings and impression sections. Through this hierarchical generation strategy, the model is able to capture the clinical logical relationship between imaging observations and diagnostic conclusions.

To optimize MedChain, we jointly consider visual evidence learning, medical semantic reasoning, and structural consistency constraints during clinical report generation. Specifically, the overall training objective consists of four components: report generation loss, clinical evidence reasoning loss, medical semantic reasoning loss, and structural consistency loss. Since the proposed framework adopts a hierarchical report generation strategy and decomposes the generation process into two stages, findings and impression, the report generation loss consists of the generation objectives of both stages. For the findings generation stage, the model produces imaging findings based on the clinical evidence representation E, and the optimization objective is defined as

$$L_f = -\sum_{t=1}^{T_f} log\, P(y_t^f \left| y_{<t}^f, E \right.) \tag{16}$$

where T_f denotes the length of the findings sequence, and y_t^f represents the t-th findings token.

For the impression generation stage, the model further incorporates the generated findings information and the medical semantic representation C_E fused with visual evidence to generate diagnostic summaries. The optimization objective is formulated as

$$L_i = -\sum_{t=1}^{T_i} log\, P(y_t^i \left| y_{<t}^i, Y_f, C_E \right.) \tag{17}$$

where T_i denotes the length of the impression sequence, and Y_f represents the generated findings content.

Accordingly, the hierarchical report generation loss is defined as

$$L_{rep} = L_f + L_i \tag{18}$$

To enhance the capability of the model to characterize the visual discrepancy between patient image states and normal reference states, we introduce a clinical evidence reasoning loss to constrain the learning process of visual evidence representations. Specifically, this loss encourages the model to focus on regions with significant deviations from normal reference representations while suppressing visual variations that are irrelevant to clinical diagnosis. Based on the deviation magnitude between visual features and normal reference representations, regions with larger deviations are regarded as potential abnormality-related evidence, whereas regions with smaller deviations are considered normal reference regions. Corresponding visual evidence representations are then constructed accordingly. Let E_p denote the visual evidence representation with larger deviation from the normal reference representation, and E_n denote the visual evidence representation with smaller deviation. The clinical evidence reasoning loss is defined as

$$L_{evi} = max(0, m - \|E_p - E_n\|_2) \tag{19}$$

where m denotes the distance margin, fixed at 1.0. The hinge penalty is active only when the Euclidean separation between higher-deviation and lower-deviation evidence representations falls below this margin and vanishes at or above the margin. These groups are defined by feature deviations from the learned normal references rather than by independent lesion annotations.

For each examination, regional deviation scores are the Euclidean distances between its visual tokens and their attention-retrieved normal references. The distances are min–max normalized within the examination, mapping the smallest value to zero and the largest to one. A normalized deviation threshold of 0.50 separates higher-deviation evidence from lower-deviation evidence. Tokens scoring at least 0.50 are assigned to the former group, and tokens scoring below 0.50 to the latter. The evidence vectors in the margin loss are formed by averaging the difference-aware attention outputs within the respective groups. If all regional distances coincide or a group is empty, the pairwise evidence loss for that examination is omitted. This feature-based grouping is distinct from image-confirmed lesion annotation.

To further constrain the model to learn semantic associations between visual evidence and medical concepts, we design a medical semantic reasoning loss. During training, medical concept supervision labels are constructed from concepts extracted from radiology reports and mapped into a predefined medical concept space. Let the ground-truth medical concept label vector be $C^* = \{c_1^*, c_2^*, \ldots, c_M^*\}$ and the predicted medical concept probability vector be $\hat{C} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_M\}$. The medical semantic reasoning loss is formulated using binary cross-entropy:

$$L_{rea} = -\sum_{j=1}^{M} \left[c_j^* \log \hat{c}_j + (1 - c_j^*) \log(1 - \hat{c}_j) \right] \tag{20}$$

where M denotes the number of medical concepts. This loss encourages the model to learn the latent associations among imaging findings, medical concepts, and disease states, thereby improving its medical reasoning capability.

Furthermore, since the Findings and Impression sections in medical reports describe imaging observations and diagnostic summaries, respectively, they should maintain case-level semantic consistency despite containing different types of information. Therefore,

we introduce a structural consistency loss to constrain the semantic relationship between the two report generation stages. Specifically, the hidden states generated during the findings and impression generation processes are independently averaged to obtain their corresponding case-level semantic representations:

$$z_f = MeanPool(H_f) \tag{21}$$

$$z_i = MeanPool(H_i) \tag{22}$$

The consistency between the two semantic representations is then measured using cosine distance:

$$L_{str} = 1 - \frac{z_f^T z_i}{\|z_f\|_2 \|z_i\|_2} \tag{23}$$

This loss does not require findings and impression to be identical at the textual level. Instead, it constrains them to remain consistent in the case-level semantic space, thereby reducing potential mismatches between imaging descriptions and diagnostic conclusions.

Overall, the optimization objective of MedChain is defined as

$$L_{total} = L_{rep} + \lambda_1 L_{evi} + \lambda_2 L_{rea} + \lambda_3 L_{str} \tag{24}$$

where λ_1, λ_2 and λ_3 are weighting coefficients used to balance the contributions of the clinical evidence reasoning loss, medical semantic reasoning loss, and structural consistency loss in the overall training objective, respectively.

4. Experiments and Discussion

4.1. Dataset

We conducted experiments on the Indiana University Chest X-ray Collection (IU X-ray) [28] of Indiana University to evaluate the effectiveness of our proposed methods. IU X-ray is a publicly available chest X-ray dataset consisting of 7470 frontal- and lateral-view chest X-ray images and their corresponding 3955 diagnostic reports. Following previous IU X-Ray-based report generation studies, we randomly split the examinations into training, validation, and test sets with a ratio of 7:1:2 [10]. Since a single examination may contain both frontal and lateral views, the dataset partition was performed at the examination level. All images and reports associated with the same examination were assigned exclusively to one subset to prevent potential information leakage caused by different views of the same examination appearing in both training and testing sets. The training set was used for model optimization and construction of the normal-state memory bank. The validation set was used for model selection and hyperparameter tuning, while the test set was reserved exclusively for final performance evaluation.

For image preprocessing, all chest X-ray images were first standardized according to the input requirements of the visual encoder, and pixel normalization was performed following the corresponding pre-training configuration. During training, lightweight data augmentation strategies were applied to improve the robustness of the model against variations in acquisition conditions while preserving the integrity of anatomical structures and pathological regions. For textual data, the original radiology reports were normalized by converting text to a unified format, removing redundant spaces and meaningless formatting symbols, and performing tokenization. A report vocabulary was constructed based on the training set, with special tokens introduced for sequence start, sequence end, unknown words, and padding. To preserve the clinical organization structure of radiology reports, we separately retained the findings and impression sections rather than treating reports as unstructured continuous text. The findings section mainly serves as supervision for imaging evidence description, while the impression section is used to learn further semantic

reasoning from imaging findings to diagnostic summaries, which is consistent with the hierarchical clinical report generation strategy proposed in this work.

Furthermore, to support normality-aware clinical evidence reasoning, we constructed a set of candidate normal cases from the training data and used them to establish the normal-state memory bank. Candidate normal cases were selected based on clinical observations in radiology reports, excluding examinations containing explicit abnormal imaging findings or disease-related diagnostic descriptions. This strategy aims to obtain imaging samples that approximate normal anatomical states for constructing normal reference representations. The normal-state memory bank was built exclusively using training cases, and samples from the validation and test sets were not involved in the construction of normal visual prototypes. Through this strategy, candidate normal cases are no longer treated merely as negative samples in traditional classification tasks. Instead, they are utilized to characterize the distribution of normal anatomical states in chest X-ray images. For a given patient image, the model retrieves the corresponding normal reference representation from the normal-state memory bank based on visual similarity and further extracts the discrepancy between the patient state and the normal state, thereby highlighting potential abnormal regions. This process is consistent with the normal-state representation and difference-aware feature modeling design in the proposed clinical evidence reasoning module.

For comparison beyond IU X-Ray, we consider MIMIC-CXR [29] and the corresponding MIMIC-CXR-JPG v2.1.0 resource [30]. MIMIC-CXR comprises 377,110 radiographs from 227,835 studies at a separate clinical institution. The JPG resource provides the official training, validation, and test split, along with view metadata. Evaluation units are radiographic studies, and all images associated with one examination remain within the same partition. Input-view availability and study identifiers are retained so that comparisons do not mix views or reports from different examinations.

4.2. Implementation Details and Metrics

All experiments were implemented using PyTorch 2.6.0 and conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory. The model was optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}. A cosine annealing strategy was adopted to gradually reduce the learning rate during training. The weight decay coefficient was set to 1×10^{-5}. The model was trained for 100 epochs, and the checkpoint achieving the best performance on the validation set was selected for final evaluation on the test set.

During training, random cropping, random brightness adjustment, and random contrast adjustment were applied as data augmentation strategies to improve model robustness against variations in acquisition conditions. Horizontal flipping was not applied because the left–right anatomical orientation of chest radiographs contains clinically meaningful information and may influence the interpretation of abnormal findings. The visual encoder adopts a Swin-Transformer architecture to extract multi-scale image representations. The visual backbone was initialized with ImageNet pre-trained parameters. The extracted visual features from different stages were projected into a unified feature space with a dimension of 512. The multi-scale visual tokens were subsequently used as inputs for abnormal evidence reasoning and medical knowledge interaction modules.

For the normality-aware clinical evidence reasoning module, candidate normal cases were selected from the training set according to the absence of abnormal findings in radiology reports. Only training samples were used to construct the normal state memory bank to avoid information leakage between training and testing phases. The reported MedChain configuration uses a normal memory bank with 64 learnable normal visual prototypes, each represented by a 512-dimensional feature vector. During inference, patient visual

representations were compared with normal prototypes to obtain normal reference features, which were then used to generate difference-aware visual evidence representations.

The medical knowledge reasoning module was constructed using a medical concept graph containing anatomical entities, radiological observations, and disease-related concepts. The graph encoder consisted of two graph Transformer layers with eight attention heads. Each node representation was mapped into the same 512-dimensional semantic space as visual features. The knowledge-enhanced representations were fused with difference-aware visual evidence through cross-modal attention.

The report generation module was implemented using a Transformer-based hierarchical decoder. The decoder consisted of three Transformer layers with eight attention heads. The hidden dimension of the decoder was set to 512, and the feed-forward layer dimension was set to 2048. Dropout was set to 0.1 in all Transformer layers. Instead of directly generating the whole report as a single sequence, the proposed model generated the Findings section first and subsequently generated the impression section conditioned on the extracted findings and medical concepts.

For textual preprocessing, all reports were converted into lowercase and tokenized using a word-level tokenizer. Words appearing fewer than three times in the training corpus were replaced with the unknown token. The final vocabulary contained approximately 4000 unique tokens. The maximum length of generated reports was limited to 100 tokens, and shorter sequences were padded with a special padding token.

During training, the batch size was set to 16. Teacher forcing was adopted for report generation, and the model was optimized using the cross-entropy loss between generated tokens and ground-truth reports. The weight coefficients of auxiliary objectives, including abnormal evidence alignment, medical knowledge consistency, and structural consistency losses, were experimentally set to 0.5, 0.1, and 0.1, respectively. The final loss was obtained by combining the report generation loss with these auxiliary constraints. During inference, beam search decoding was employed with a beam size of 3.

All baseline models were evaluated using the same IU X-Ray data split and preprocessing pipeline whenever publicly available implementations were provided. We report results from original papers when official implementations were unavailable, while noting that these results may be affected by differences in experimental settings. To comprehensively assess the performance of MedChain in X-ray report generation, we employ a series of standard metrics widely used in medical captioning tasks. Specifically, we report the bilingual evaluation understudy (BLEU-n) [31] to measure n-gram precision and overlap between generated reports and reference texts. In addition, we adopt the Metric for Evaluation of Translation with Explicit ORdering (METEOR) [32] to account for synonym matching and morphological alignment, alongside the longest common subsequence-based F-measure of recall-oriented understudy for gisting evaluation (ROUGE_L) [33] to evaluate structural similarity and sentence-level recall. These metrics quantify agreement with reference wording, whereas clinical-observation agreement is evaluated separately using CheXbert.

The LLM triple extractor is evaluated against manually adjudicated reference triples from radiology reports in the graph-construction corpus. Predicted and reference triples are compared after consistent terminology normalization as directed source–relation–target tuples. A prediction is counted as a true positive only when both medical entities, the relation label, and edge direction match the reference. Unmatched predictions and reference triples count as false positives and false negatives, respectively. The pooled counts define micro-precision, micro-recall, and micro-F1, while relation-specific counts evaluate modify, located_at, and suggestive_of separately. The analysis also distinguishes spurious and omitted relations, incorrect anatomical localization, and disease associations affected by negation or uncertainty.

Clinical efficacy is evaluated by applying the same fixed CheXbert report labeler [34] to each generated report and its reference report. CheXbert recognizes 14 clinical observations, including No Finding. The clinical observation categories include Cardiomegaly, Edema, Atelectasis, Pneumothorax, and Pleural Effusion. Only a Positive output is counted as present. Negative, Uncertain, and Blank outputs are treated as non-positive. Micro-precision, micro-recall, and micro-F1 are calculated from true-positive, false-positive, and false-negative counts pooled across examinations and observation categories. The No Finding category is included in the pooled score and distinguished from the remaining observations when interpreting errors. These metrics quantify concordance between report-derived clinical labels rather than image-confirmed diagnostic accuracy. False negatives represent reference-positive observations absent from generated reports, whereas false positives represent predicted-positive observations absent from reference reports. Both error types are distinct from n-gram mismatch.

4.3. Performance Comparison and Discussion

We assess report generation using complementary linguistic and clinical endpoints. Table 1 compares text-overlap metrics on IU X-Ray, while Table 2 reports CheXbert-based clinical observation precision, recall, and F1 on IU X-Ray and MIMIC-CXR.

Table 1. Text-overlap performance comparison of representative medical image report generation methods on the IU X-Ray dataset. The best results are highlighted in bold. The second-best values for each metric are underlined. * denotes p-value < 0.01 on a two-tailed t-test against the best baseline.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
HRGR [35]	0.438	0.298	0.208	0.151	-	0.312
R2Gen [10]	0.470	0.304	0.219	0.165	0.187	0.371
CA [7]	0.492	0.314	0.222	0.169	0.193	0.381
CMN [36]	0.475	0.309	0.222	0.170	0.191	0.375
mDiTag [11]	0.329	0.198	0.135	0.094	0.194	0.273
PGT [37]	0.482	0.313	0.232	0.181	0.203	0.381
GSKET [38]	0.496	0.327	0.238	0.178	-	0.381
PureT [39]	0.496	0.319	0.241	0.175	-	0.377
CECL [40]	0.485	0.331	0.241	0.181	<u>0.208</u>	**0.406**
MMTN [41]	0.486	0.321	0.232	0.175	-	0.361
S3-Net [42]	<u>0.499</u>	0.334	0.246	0.172	0.206	0.401
CmEAA [43]	0.481	0.319	0.234	0.181	-	0.392
RMPT [17]	0.497	0.349	0.257	0.194	0.202	0.395
FedMRG [44]	0.414	0.321	0.233	0.181	0.192	0.275
KAM [45]	0.494	<u>0.353</u>	<u>0.264</u>	<u>0.201</u>	<u>0.208</u>	0.378
MedChain	**0.523 ***	**0.362 ***	**0.281 ***	**0.206 ***	**0.213 ***	<u>0.405</u>

Table 2. CheXbert-based clinical efficacy results across IU X-Ray and MIMIC-CXR evaluation settings. Precision, recall, and F1 are micro-averaged over the 14 CheXbert report-label categories, including No Finding, using the pooled TP, FP, and FN counts.

Model and Test Setting	TP	FP	FN	CE Precision	CE Recall	CE F1
IU-trained MedChain on IU	532	822	544	0.393	0.494	0.438
IU-trained MedChain on MIMIC	1494	3897	3192	0.277	0.319	0.297
MIMIC-trained w/o MRG	2085	3058	2601	0.405	0.445	0.424
MIMIC-trained MedChain	2109	3160	2577	0.400	0.450	0.424

Table 1 shows that MedChain leads the compared methods on BLEU-1 through BLEU-4 and METEOR. Its BLEU-3 and BLEU-4 scores of 0.281 and 0.206 exceed the KAM scores of 0.264 and 0.201 by 6.44% and 2.49% relative, while METEOR reaches 0.213 against the next-best 0.208. ROUGE_L is 0.405, slightly below CECL at 0.406. These measures characterize agreement with reference wording rather than the recovery of individual positive clinical observations. The effects of difference-aware visual evidence, graph reasoning, and hierarchical decoding are examined separately by the component ablations in Table 3.

Table 3. Ablation study of different components in MedChain on IU X-Ray. The best results are highlighted in bold.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
w/o CERM	0.509	0.346	0.263	0.191	0.205	0.394
w/o MRG	0.514	0.351	0.271	0.198	0.202	0.399
w/o HCRG	0.517	0.355	0.274	0.201	0.208	0.395
MedChain	**0.523**	**0.362**	**0.281**	**0.206**	**0.213**	0.405

For clinical-label assessment, MIMIC-CXR provides a larger external benchmark drawn from a different clinical institution. We evaluate two complementary configurations to distinguish cross-institution transfer from target-domain training. For transfer, the IU-trained MedChain checkpoint is evaluated on the official MIMIC-CXR test studies with its visual encoder, normal-reference memory, reasoning graph, vocabulary, and decoder frozen. For target-domain training, MedChain and its variant without MRG are trained on the official MIMIC-CXR training split, with checkpoint selection performed on the validation split. The evaluation uses consistent study-level image selection, preprocessing, decoding, and clinical-label scoring within each comparison.

Table 2 reports clinical-label agreement using pooled CheXbert counts. IU-trained MedChain attains CE precision of 0.393, recall of 0.494, and F1 of 0.438 on IU X-Ray. When the same checkpoint is transferred without adaptation to MIMIC-CXR, precision decreases to 0.277, recall to 0.319, and F1 to 0.297. The absolute F1 difference of 0.141 reflects lower agreement with reference-positive findings across institutions. Training MedChain on MIMIC-CXR yields an F1 of 0.424 on its test set, an absolute increase of 0.127 over the frozen IU-trained configuration. On the same MIMIC-CXR test distribution, target-domain training increases precision by 0.123 and recall by 0.131 relative to frozen transfer.

On the MIMIC-CXR test set, target-domain training increases correctly matched positive observations from 1494 to 2109 and decreases missed reference-positive observations from 3192 to 2577. False-positive clinical labels simultaneously decrease from 3897 to 3160. Thus, the 0.127 F1 gain over frozen transfer reflects both greater coverage of reference-positive findings and fewer predicted-positive labels absent from the reference. The two evaluation configurations distinguish the limitations of unadapted cross-institution transfer from MedChain's capacity to learn on a larger, independent chest-radiograph corpus.

The MIMIC-CXR-trained graph ablation tests whether structured relations improve clinical-observation agreement on the same test studies. Complete MedChain yields CE precision of 0.400 and recall of 0.450, compared with 0.405 and 0.445 without MRG. It recovers 24 additional reference-positive labels but also produces 102 additional false-positive labels, leaving both variants at CE F1 of 0.424 to three decimals. The recall increase therefore trades against precision rather than improving aggregate F1. Together with the positive IU X-Ray n-gram ablation, this comparison shows that better wording agreement with graph reasoning does not automatically imply better clinical-label agreement.

4.4. Case Study

Figure 3 compares the generated reports, radiologist-written reports, and reference reports for three IU X-Ray examinations. Cases 1 and 3 contain multiple reported abnormalities, whereas Case 2 represents a normal examination. The comparison focuses on preserved findings, omitted observations, and additional diagnostic statements. Figure 3 additionally visualizes heatmaps, where warmer colors indicate stronger abnormality-related responses.

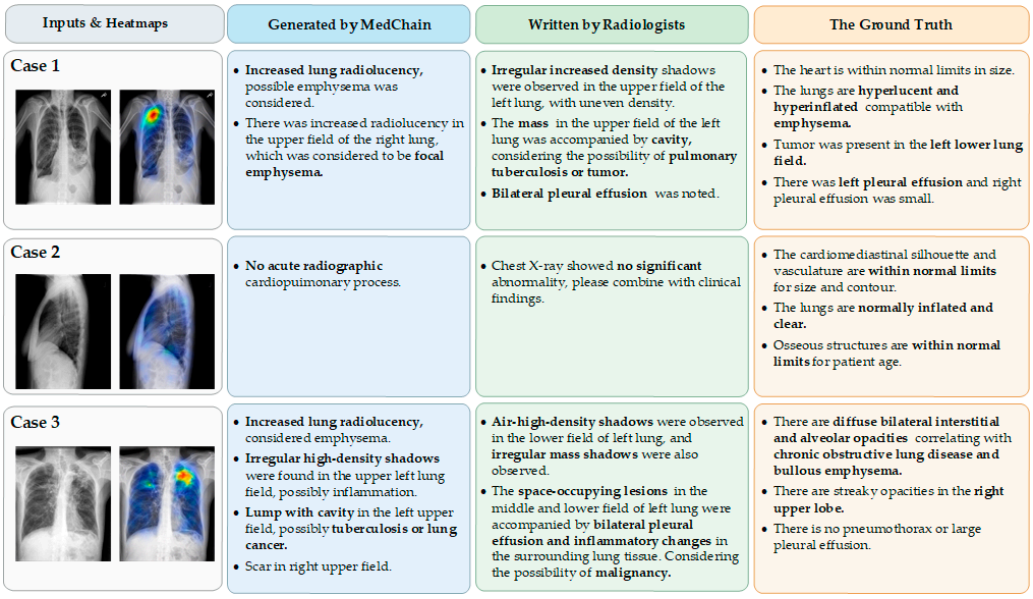

Figure 3. Qualitative comparison of generated reports on representative IU X-Ray cases with CERM evidence heatmaps. The first column shows the input image and a superimposed heatmap, followed by the generated report, radiologist-written report, and ground-truth report. Warmer colors indicate stronger abnormality-related evidence responses. Clinically relevant findings and diagnostic concepts are highlighted in bold.

In Case 1, MedChain describes increased lung lucency and possible emphysema, consistent with the reference description of hyperlucent and hyperinflated lungs. The reference also records a tumor in the left lower lung field and bilateral pleural effusions, but neither abnormality appears in the generated report. The additional description of focal emphysema in the right upper lung field does not recover these omitted findings. This case illustrates the incomplete coverage of coexisting abnormalities despite the recognition of an emphysema-related pattern.

In Case 2, MedChain generates a normal cardiopulmonary assessment consistent with the radiologist-written report and the reference description of normal lung and mediastinal findings. Its report is less detailed than the reference but does not introduce a positive disease assertion in this case.

In Case 3, the generated report includes increased lung lucency, a left upper-lung opacity described as a cavitary mass, right upper-lung scarring, and differential diagnoses of inflammation, tuberculosis, and lung cancer. The reference instead describes bilateral diffuse interstitial and alveolar opacities associated with chronic obstructive pulmonary disease and bullous emphysema, together with right upper-lobe linear opacities. The generated report therefore does not preserve the bilateral distribution and full extent of the reference-described parenchymal abnormalities.

The left upper-lung cavitary mass and its associated disease hypotheses are absent from the reference report. These discrepancies identify potentially unsupported content, but establishing that the statements lack imaging evidence would require independent

review of the radiograph. The generated report also omits the reference statement that no pneumothorax or large pleural effusion is present. That missing negative observation should not be counted as an omitted lesion.

Cases 1 and 3 reveal distinct errors in multi-abnormality reporting. Missing abnormality-location descriptions indicate incomplete finding coverage, whereas disease hypotheses not corroborated by the reference observations raise a separate question of evidential support. These cases motivate a finding-level audit that distinguishes omitted abnormalities, mismatched anatomical locations, and visually unsupported diagnostic claims rather than inferring any of these rates from text-overlap scores.

To mitigate omission of coexisting abnormalities, a possible extension is to retain multiple spatially diverse CERM evidence candidates and align them with anatomy-aware finding slots derived from the medical reasoning graph before Findings decoding. A coverage constraint on slot attention could discourage repeated selection of the most salient pattern, while a learned relevance gate would avoid treating every high-residual region as pathological. This design targets the failure to report the reference-described tumor and pleural effusions in Case 1. To mitigate unsupported disease inferences, an evidence-gated Impression stage could associate each proposed disease concept with an explicit Findings observation and its corresponding visual evidence. When this patient-specific support is insufficient, the decoder could retain the observable radiographic description without asserting a more specific disease hypothesis. Evaluation of these strategies should measure abnormality coverage, anatomical agreement, and image-confirmed unsupported positive statements in single- and multi-abnormality cases, since stronger gating might also suppress subtle true findings.

4.5. Ablation Study

To evaluate the contribution of each key component in MedChain to medical image report generation performance, we conduct ablation experiments on the IU X-Ray dataset. Specifically, w/o CERM denotes the variant without the clinical evidence reasoning module, where the model no longer utilizes the discrepancy between patient images and normal anatomical reference states for abnormal visual evidence modeling. w/o MRG represents the variant without the medical reasoning graph module, where the relationships among visual evidence, medical concepts, and disease states are no longer established through structured medical knowledge. w/o HCRG denotes the variant without the hierarchical clinical report generation strategy, where the original two-stage generation process from findings to impression is replaced by directly generating the complete report sequence.

All ablation variants adopt the same data split, training strategy, and evaluation metrics as MedChain to ensure that the experimental results accurately reflect the actual contribution of each component to model performance. Table 3 presents the performance of different model variants on 6 evaluation metrics, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE_L.

As shown in Table 3, the complete MedChain achieves the best performance on all 6 evaluation metrics. Removing any key component leads to performance degradation to varying degrees, demonstrating that clinical evidence reasoning, medical knowledge reasoning, and hierarchical report generation all contribute positively to the final report quality. After removing the clinical evidence reasoning module, the w/o CERM variant shows a decrease in BLEU-4 from 0.206 achieved by MedChain to 0.191, corresponding to a performance drop of approximately 7.28%. This indicates that without clinical evidence reasoning, the model's ability to generate continuous medical descriptions and fine-grained abnormality expressions is substantially affected. This phenomenon mainly results from the

fact that conventional visual features typically contain both normal anatomical structures and abnormal regions, making it difficult to distinguish diagnostically meaningful visual variations. CERM constructs discrepancy representations between patient image states and normal anatomical reference states, enabling the model to focus on regions deviating from normal anatomical patterns and obtain more clinically meaningful visual evidence representations. Without this module, the model lacks an explicit abnormal evidence modeling process, making the generated reports more susceptible to irrelevant visual information and resulting in reduced accuracy of medical descriptions.

Removing the medical reasoning graph also reduces report generation performance. Compared with MedChain, w/o MRG decreases BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by 1.72%, 3.04%, 3.56%, and 3.88%, respectively, while the METEOR score decreases by 5.16%. These results demonstrate that medical knowledge reasoning plays an important role in improving the semantic consistency of generated reports. Visual manifestations in medical images usually cannot be directly mapped to unique disease concepts. For example, the same imaging finding may correspond to different disease states, and the same disease may present with various imaging characteristics. MRG constructs a medical relational network containing anatomical structures, imaging findings, and disease concepts, providing a structured semantic reasoning space for visual evidence and enabling the model to learn associations between visual abnormalities and clinical concepts. After removing this module, the model must directly map visual representations to textual descriptions without medical knowledge constraints, making it more likely to generate reports with insufficient medical semantics or incomplete concept associations. This also explains the relatively larger degradation of w/o MRG on the METEOR metric, as METEOR emphasizes semantic similarity rather than relying solely on lexical overlap.

Furthermore, removing the hierarchical clinical report generation strategy results in relatively smaller performance degradation but still shows a consistent downward trend. The w/o HCRG variant decreases BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by 1.15%, 1.93%, 2.49%, and 2.43%, respectively. Meanwhile, METEOR and ROUGE_L decrease by 2.35% and 2.47%, respectively. These results indicate that without decomposing report generation into two stages, namely findings generation and impression generation, and without applying structural consistency constraints, the model becomes less effective in maintaining long-range textual structure, sentence-level similarity, and contextual logical coherence. In real clinical reporting workflows, radiologists usually describe imaging findings first and then formulate diagnostic impressions based on these observations. The hierarchical generation strategy proposed in this work simulates this process. After removing this strategy, the model is required to directly generate the complete report sequence without an intermediate semantic organization stage, leading to reduced structural consistency and weaker long-range semantic dependency modeling.

Overall, the three ablation experiments validate the effectiveness of each component in MedChain. Specifically, CERM mainly enhances the model's ability to identify clinically meaningful abnormal evidence, MRG provides knowledge-based reasoning capability from visual evidence to medical concepts, and HCRG improves clinical structural consistency during report generation. By integrating abnormal evidence modeling, medical knowledge reasoning, and hierarchical report generation, the complete model establishes a comprehensive process from image observation to medical semantic understanding and clinical text generation. The consistent superiority of MedChain over all ablation variants across evaluation metrics further demonstrates the effectiveness of the proposed structured clinical reasoning framework for medical image report generation.

4.6. Hyperparameter Analysis of Normal-Reference Prototype Capacity

The number of normal-reference prototypes determines the capacity of the reference memory. We consider K values of 16, 32, 64, 128, and 256 with a fixed feature dimension of 512. Table 4 presents the corresponding report-generation metrics.

Table 4. Report-generation metrics across different numbers of normal-reference prototypes. The highest values in each column are highlighted in bold.

K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
16	0.521	0.359	0.277	0.203	0.212	0.402
32	0.522	0.361	0.279	0.204	0.212	0.403
64	**0.523**	**0.362**	**0.281**	**0.206**	**0.213**	**0.405**
128	0.522	0.361	0.280	**0.206**	**0.213**	0.404
256	0.522	0.360	0.280	**0.206**	0.212	0.404

Each prototype contains 512 learnable parameters. The total number of prototype parameters therefore increases from 8192 at K = 16 to 131,072 at K = 256. The number of prototype comparisons per visual token also grows with K.

Memory capacity affects the granularity of available normal reference patterns, while larger banks require more similarity computations. Selecting K therefore involves balancing the coverage of normal visual patterns against the additional retrieval cost.

4.7. LLM Triple-Extraction Accuracy Evaluation

The independent extraction audit uses study-level reference annotations from the training corpus employed to construct the medical reasoning graph. Its annotation schema includes anatomical and observation entities, negation and uncertainty status, and directed modify, located_at, and suggestive_of relations. After terminology normalization, extracted triples are matched to the adjudicated references by entity pair, relation label, and edge direction. Report-level bootstrap resampling provides an uncertainty interval for the pooled F1 score. Scoring extracted edges before graph encoding directly evaluates the reliability of the graph-construction stage.

Table 5 reports an overall micro-precision of 0.756, micro-recall of 0.781, and micro-F1 of 0.768 from 813 true-positive, 263 false-positive, and 228 false-negative triples. The relation-level F1 scores are 0.768 for modify, 0.781 for located_at, and 0.737 for suggestive_of, yielding a macro-F1 of 0.762. Relation quality is uneven. The suggestive_of triples have precision 0.717 and F1 0.737, compared with 0.767 and 0.781 for located_at. Their 58 false-positive edges can introduce incorrect finding-to-disease links into graph propagation, whereas 47 missed edges remove potentially relevant links. The measured error profile identifies disease-association extraction as the weakest part of the graph input in this audit. The report-level bootstrap 95% confidence interval for the overall micro-F1 is 0.748 to 0.788.

Table 5. Precision, recall, and F1 of LLM-extracted medical triples by relation type.

Relation or Summary	TP	FP	FN	Precision	Recall	F1
modify	284	89	83	0.761	0.774	0.768
located_at	382	116	98	0.767	0.796	0.781
suggestive_of	147	58	47	0.717	0.758	0.737
Overall micro-average	813	263	228	0.756	0.781	0.768

5. Conclusions

Existing medical image report generation methods typically formulate the task as a direct mapping from visual features to textual descriptions, which overlooks the clinical reasoning process involved in radiologists' decision making. To address this limitation, we propose MedChain, a knowledge-guided evidence reasoning framework that integrates normality-aware visual evidence extraction, medical knowledge graph reasoning, and hierarchical report generation. By modeling deviations from normal anatomical reference representations, MedChain extracts difference-aware visual evidence and establishes structured associations among imaging findings, anatomical structures, and disease concepts. The hierarchical generation strategy further organizes this information into findings and diagnostic impressions, providing a structured pathway from visual observation to clinical reporting. Experiments on IU X-Ray demonstrate improvements in report generation performance across multiple linguistic evaluation metrics, while CheXbert-based evaluation provides complementary evidence of clinical observation agreement. Results on MIMIC-CXR further reveal the influence of cross-institutional domain differences and the benefits of target-domain training, although improvements in linguistic metrics do not necessarily translate into higher clinical-label agreement. Despite these findings, several challenges remain. In examinations containing multiple abnormalities, MedChain may fail to capture all clinically relevant findings or accurately preserve their anatomical locations and spatial distributions. Furthermore, the propagation of imperfect medical knowledge relations may introduce disease associations that are insufficiently supported by patient-specific imaging evidence. These limitations highlight the need for more comprehensive lesion-level evidence modeling and stronger alignment between visual observations and diagnostic reasoning. Future work will therefore focus on improving multi-abnormality coverage, anatomical localization, and evidence-grounded impression generation to enhance the clinical reliability of automated radiology reports.

Author Contributions: Conceptualization, J.L.; Methodology, K.C.; Software, K.C.; Validation, W.W., Z.D.; Writing—original draft, K.C.; Writing—review and editing, W.W.; Visualization, H.W.; Supervision, Z.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jimei University, grant number ZQ2026034.

Institutional Review Board Statement: The Institutional Review Board (IRB) approval was waived because this study involved only secondary analysis of previously collected, publicly available, de-identified datasets (IU X-Ray and MIMIC-CXR), without any new clinical data collection or direct involvement of human participants.

Informed Consent Statement: Informed consent was waived because the study exclusively used previously collected, de-identified data from these publicly available datasets, and no direct interaction with patients or access to identifiable personal information was involved.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We acknowledge the use of GPT-6.1 Sol solely to assist in formatting the manuscript in accordance with the formatting requirements and guidelines.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, D.; Kong, M.; Zhao, Y.; Huang, J.; Huang, Z.; Kuang, K.; Wu, F.; Zhu, Q. Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning. *Med. Image Anal.* **2024**, *91*, 102982. [CrossRef] [PubMed]
2. Lee, S.; Youn, J.; Kim, H.; Kim, M.; Yoon, S.H. CXR-LLaVA: A multimodal large language model for interpreting chest X-ray images. *Eur. Radiol.* **2025**, *35*, 4374–4386. [CrossRef] [PubMed]

3. Wang, L.; Chen, H.; Liu, Y.; Lyu, Y.; Qiu, F. Reinforced visual interaction fusion radiology report generation. *Multimed. Syst.* **2024**, *30*, 299. [CrossRef]
4. Farkh, R.; Oudinet, G.; Foued, Y. Image Captioning Using Multimodal Deep Learning Approach. *Comput. Mater. Contin.* **2024**, *81*, 3951–3968. [CrossRef]
5. Sun, S.; Su, Z.; Meizhou, J.; Feng, Y.; Hu, Q.; Luo, J.; Hu, K.; Yang, Z. Optimizing medical image report generation through a discrete diffusion framework. *J. Supercomput.* **2025**, *81*, 637. [CrossRef]
6. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
7. Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; Sun, X. Contrastive Attention for Automatic Chest X-ray Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 269–280.
8. Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S.K.; Xiao, L. Radiology report generation with a learned knowledge base and multi-modal alignment. *Med. Image Anal.* **2023**, *86*, 102798. [CrossRef] [PubMed]
9. You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; Wu, X. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 72–82.
10. Chen, Z.; Song, Y.; Chang, T.H.; Wan, X. Generating Radiology Reports via Memory-driven Transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 1439–1449.
11. Park, H.; Kim, K.; Yoon, J.; Park, S.; Choi, J. Feature Difference Makes Sense: A Medical Image Captioning Model Exploiting Feature Difference and Tag Information. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 5–10 July 2020; pp. 95–102.
12. Xu, D.; Zhu, H.; Huang, Y.; Jin, Z.; Ding, W.; Li, H.; Ran, M. Vision-knowledge fusion model for multi-domain medical report generation. *Inf. Fusion* **2023**, *97*, 101817. [CrossRef]
13. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2577–2586.
14. Peng, S.; Xiong, C.; Liu, L.; Yang, L.T.; Chen, J. GRPIC: An end-to-end image captioning model using three visual features. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 1559–1572. [CrossRef]
15. Shafiq, H.; Gilanie, G.; Sajid, M.; Ahsan, M. Dental radiology: A convolutional neural network-based approach to detect dental disorders from dental images in a real-time environment. *Multimed. Syst.* **2023**, *29*, 3179–3191. [CrossRef]
16. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
17. Qin, C.; Xiong, Y.; Chen, W.; Li, Y. RMPT: Reinforced Memory-Driven Pure Transformer for Automatic Chest X-Ray Report Generation. *Mathematics* **2025**, *13*, 1492. [CrossRef]
18. Shi, J.; Wang, S.; Wang, R.; Ma, S. Aimnet: Adaptive image-tag merging network for automatic medical report generation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7737–7741.
19. Wang, Z.; Liu, L.; Wang, L.; Zhou, L. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 18–22 June 2023; pp. 11558–11567.
20. Liu, Y.; Wang, C.; Lu, M.; Yang, J.; Gui, J.; Zhang, S. From Simple to Complex Scenes: Learning Robust Feature Representations for Accurate Human Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5449–5462. [CrossRef] [PubMed]
21. Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; Xu, D. When Radiology Report Generation Meets Knowledge Graph. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
22. Huang, Z.; Zhang, X.; Zhang, S. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19809–19818.
23. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Proceedings of the Findings of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022*; ACL: Stroudsburg, PA, USA, 2022; pp. 448–458.
24. Wang, S.; Wang, C.; Shi, C.; Liu, Y.; Lu, M. Mask-Guided Mamba Fusion for Drone-Based Visible-Infrared Vehicle Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5005712. [CrossRef]
25. Hou, W.; Xu, K.; Cheng, Y.; Li, W.; Liu, J. ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1, pp. 8108–8122.
26. Yi, X.; Fu, Y.; Liu, R.; Zhang, H.; Hua, R. TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 2152–2162. [CrossRef]

27. Qwen Team. Qwen3.5-9B: Model Card. Hugging Face. 2026. Available online: https://huggingface.co/Qwen/Qwen3.5-9B (accessed on 5 November 2026).
28. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.K.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]
29. Johnson, A.E.W.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
30. Johnson, A.; Lungren, M.; Peng, Y.; Lu, Z.; Mark, R.; Berkowitz, S.; Horng, S. *MIMIC-CXR-JPG—Chest Radiographs with Structured Labels; Version 2.1.0*; PhysioNet: Cambridge, MA, USA, 2024. [CrossRef]
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
32. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
33. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS), Barcelona, Spain, 25–26 July 2004; pp. 74–81.
34. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.; Lungren, M. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 1500–1519. [CrossRef]
35. Li, C.Y.; Liang, X.; Hu, Z.; Xing, E.P. Hybrid retrieval-generation reinforced agent for Medical Image Report Generation. In *Advances in Neural Information Processing Systems, Proceedings of the32nd Annual Conference on Neural Information Processing Systems, NeurIPS 2018, Montreal, QC, Canada, 2–8 December 2018*; Curran Associates Inc.: Red Hook, NY, USA; pp. 1537–1547.
36. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal Memory Networks for Radiology Report Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 5904–5914.
37. Yan, B.; Pei, M.; Zhao, M.; Shan, C.; Tian, Z. Prior Guided Transformer for Accurate Radiology Reports Generation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5631–5640. [CrossRef] [PubMed]
38. Yang, S.; Wu, X.; Ge, S.; Zhou, S.K.; Xiao, L. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* **2022**, *80*, 102510. [CrossRef] [PubMed]
39. Wang, Z.; Han, H.; Wang, L.; Li, X.; Zhou, L. Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach. *IEEE Trans. Med. Imaging* **2022**, *41*, 2803–2813. [CrossRef] [PubMed]
40. Liu, X.; Xin, J.; Shen, Q.; Li, C.; Huang, Z.; Wang, Z. End-to-End Clustering Enhanced Contrastive Learning for Radiology Reports Generation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *9*, 1780–1794. [CrossRef]
41. Cao, Y.; Cui, L.; Zhang, L.; Yu, F.; Li, Z.; Xu, Y. MMTN: Multi-Modal Memory Transformer Network for Image-Report Consistent Medical Report Generation. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 277–285. [CrossRef]
42. Pan, R.; Ran, R.; Hu, W.; Zhang, W.; Qin, Q.; Cui, S. S3-Net: A Self-Supervised Dual-Stream Network for Radiology Report Generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1448–1459. [CrossRef] [PubMed]
43. Huang, X.; Han, Y.; Li, R.; Wu, P.; Zhang, K. CmEAA: Cross-modal enhancement and alignment adapter for radiology report generation. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 9–24 January 2025; pp. 8546–8556.
44. Metmer, H.; Yang, X. FedMRG: Federated Medical Report Generation via Text-Aware Learning Rate Adjustment and Multi-Level Prototype Collaboration. *Multimed. Syst.* **2025**, *31*, 170. [CrossRef]
45. Yang, S.; Tan, H. Knowledge-Augmented Adaptive Mechanism for Radiology Report Generation. *Mathematics* **2026**, *14*, 173. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.