TempSem-GraphNet: Temporal-Semantic Graph Network for Coherent Chest X-ray Report Generation

Zexin Mao
Southwest Forestry University

Abstract. Automated chest X-ray (CXR) report generation often overlooks the vital temporal dimension and fails to integrate diverse information modalities effectively, resulting in reports lacking coherence and clinical utility. To address these limitations, we propose TempSem-GraphNet, a novel framework that explicitly models multi-modal temporal-semantic relationships for generating coherent and accurate reports from multi-temporal CXR images. Our core innovation is the Multi-modal Temporal-Semantic Graph (TempSem-Graph), which unifies visual lesions from current and historical CXRs with semantic concepts extracted from prior reports, linked by temporal, semantic, and modality-specific edges. This graph is processed by a Hierarchical Temporal-Semantic Graph Attention Network (HTGAT) to aggregate context-rich features, which then condition a fine-tuned Large Language Model (LLM) for report generation. Evaluated on the MIMIC-CXR-JPG dataset, TempSem-GraphNet significantly outperforms state-of-the-art baselines across natural language generation and clinical entity metrics. Human evaluations further corroborate our quantitative findings, demonstrating superior temporal coherence, clinical accuracy, and utility. Our work represents a significant step towards automating longitudinal medical reporting with enhanced precision and clinical relevance.

Keywords: CXR Report Generation, Temporal Modeling, Multi-modal, Graph Neural Networks, Large Language Models

1. Introduction

Medical imaging plays a pivotal role in clinical diagnosis and patient management. Among various imaging modalities, chest X-rays (CXRs) are one of the most frequently performed examinations, providing crucial insights into a wide range of cardiopulmonary conditions. The accurate and timely interpretation of CXRs, typically documented in natural language radiology reports, is fundamental for clinical decision-making. Enhancing model interpretability and robustness, as explored in domains like face anti-spoofing [1], is crucial for building trust in such critical AI applications. However, manually generating these reports is a time-consuming and labor-intensive process for radiologists, often prone to variability in terminology and completeness [2]. This has motivated extensive research into automated CXR report generation, aiming to improve efficiency, consistency, and diagnostic quality [2].

Despite significant advancements in deep learning-based approaches for medical image captioning, current methods primarily focus on generating reports from single CXR images, often overlooking the critical temporal dimension inherent in patient care. In clinical practice, patients frequently undergo serial CXR examinations, and radiologists routinely compare current images with historical ones to track disease progression, regression, or stability [3]. Existing models struggle to capture these intricate temporal relationships and the evolution of lesions

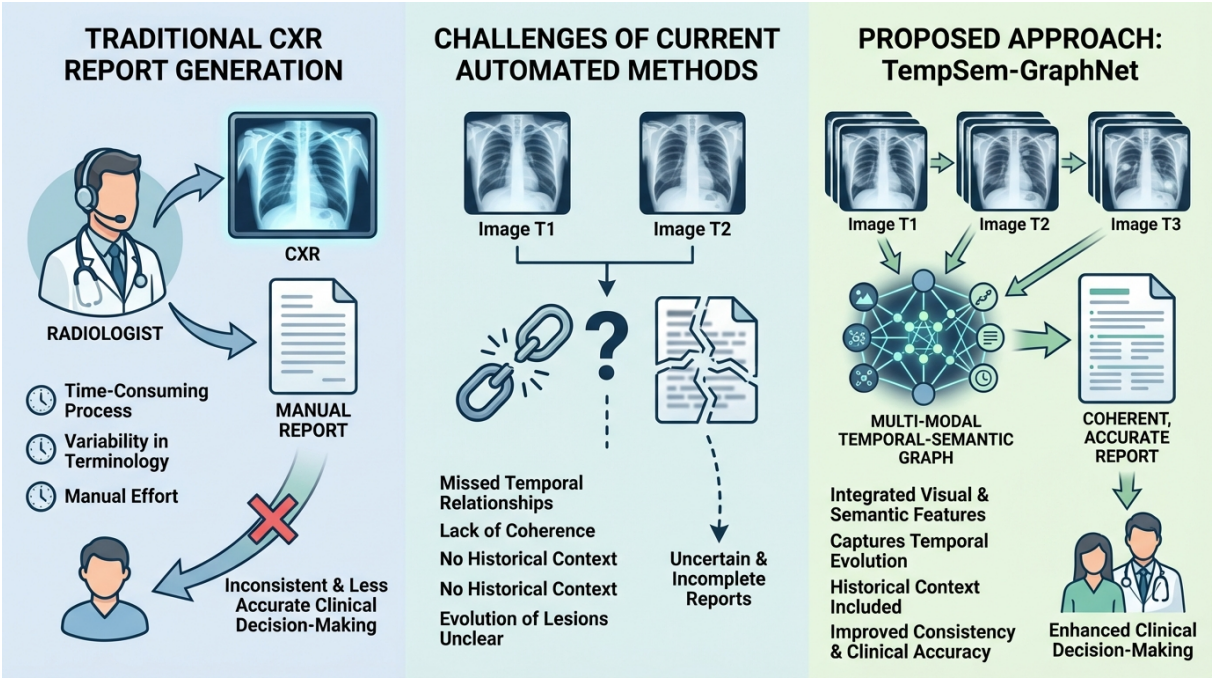


Figure 1: Overview of CXR report generation. The left panel illustrates the traditional manual process performed by radiologists, which is time-consuming and prone to inconsistencies, potentially leading to less accurate clinical decision-making. The middle panel highlights the challenges faced by current automated methods, which often fail to leverage temporal information from serial CXRs, resulting in missed temporal relationships, lack of coherence, and uncertain or incomplete reports. The right panel presents our proposed TempSem-GraphNet framework, which addresses these limitations by explicitly modeling multi-modal temporal-semantic relationships through a unified graph to generate coherent and accurate reports, thereby enhancing clinical decision-making.

over time, often producing reports that lack coherence across different time points or fail to accurately describe dynamic changes in pathology. Furthermore, effectively integrating diverse information modalities—such as visual features of lesions, their semantic descriptions, and historical context—into a unified representation remains a significant challenge. This deficiency can lead to reports that are less comprehensive and clinically less useful, especially when describing complex conditions like pneumonia resolution or tumor growth, underscoring the broader goal of multimodal agent intelligence in healthcare [4]. The pursuit of personalized and controllable content generation, as seen in recent advances in video synthesis and text-to-image models [5, 6], also highlights the increasing demand for models that can tailor outputs to specific user or contextual needs.

To address these limitations, we propose **TempSem-GraphNet: Temporal-Semantic Graph Network for Coherent Chest X-ray Report Generation**. Our method introduces a novel framework that explicitly models multi-modal temporal-semantic relationships to generate highly coherent and clinically accurate reports from multi-temporal CXR images. The core innovation lies in the construction of a *Multi-modal Temporal-Semantic Graph* (TempSem-Graph), which serves as a rich, unified representation of disease evolution. This graph integrates visual features of lesions from both current and historical CXR images with semantic concepts extracted from prior reports. Nodes in our TempSem-Graph represent individual visual lesions and semantic entities, while edges encode crucial relationships: temporal associations between

evolving lesions, semantic links between related concepts, and modality associations connecting visual findings to their textual descriptions.

The process begins by employing a frozen 2D pre-trained Vision Transformer (ViT) [7] to extract global and localized visual features from both current and historical CXR images. Concurrently, a pre-trained lesion detection model (e.g., based on YOLOv7 or RetinaNet [8]) identifies specific pathological regions, and keyword extraction techniques are applied to historical reports to distill key semantic concepts. These extracted features and concepts form the basis for constructing the TempSem-Graph. A subsequent Hierarchical Temporal-Semantic Graph Attention Network (HTGAT) then processes this graph, leveraging multi-layer message passing to effectively aggregate visual, temporal, and semantic information, thereby capturing the nuanced evolution and interdependencies of lesions. Finally, the aggregated, context-rich features from the HTGAT are fed into a fine-tuned Large Language Model (LLM), specifically LLaMA2-7B [9], which acts as a powerful decoder to generate the final natural language radiology report. Our LLM, leveraging capabilities akin to those for weak-to-strong generalization [10] and unraveling complex contexts [11], is efficiently fine-tuned using LoRA [12], ensuring high performance with minimal computational overhead.

We conducted extensive experiments on the widely used MIMIC-CXR-JPG dataset [13], a large-scale collection of CXR images and corresponding radiology reports, which includes numerous serial studies for individual patients. The CheXpert dataset [14] was additionally utilized for visual encoder pre-training and lesion detector development. Our evaluation employed both standard Natural Language Generation (NLG) metrics (BLEU-1/2/3/4, ROUGE-L, METEOR) to assess the fluency and accuracy of the generated reports, and Clinical Entity (CE) metrics (Precision, Recall, F1) to measure the clinical relevance and factual correctness of extracted medical entities. The experimental results unequivocally demonstrate that TempSem-GraphNet significantly outperforms existing baseline methods across all evaluated metrics. For instance, our method achieved a BLEU-4 score of **0.2441** and an F1-score for clinical entity extraction of **0.288**, marking substantial improvements over the best competing approaches. This superior performance highlights our model's enhanced ability to generate coherent reports that accurately reflect lesion evolution and provide precise clinical information, thanks to the explicit modeling of temporal and semantic relationships.

In summary, our main contributions are:

- We propose **TempSem-GraphNet**, a novel framework for multi-temporal CXR report generation that introduces a *Multi-modal Temporal-Semantic Graph* (TempSem-Graph) to explicitly model the evolution of lesions and integrate visual and semantic information.
- We develop a **Hierarchical Temporal-Semantic Graph Attention Network (HTGAT)** specifically designed to effectively aggregate multi-modal, temporal, and semantic information within the TempSem-Graph, enabling a comprehensive understanding of disease progression.
- We demonstrate state-of-the-art performance on the MIMIC-CXR-JPG dataset, achieving significant improvements in both natural language generation quality and the accuracy of clinical entity descriptions, thereby generating more coherent and clinically valuable radiology reports.

2. Related Work

2.1. Large Language Models and Vision-Language Models for Medical Report Generation

LLMs and VLMs are increasingly used for medical report generation from imaging data, interpreting visual information into clinical language. This section covers LLM techniques, multimodal integration, and deployment.

LLMs show potential in medical text processing and generation, e.g., GPT3Mix for data augmentation [15] and MS^2 for multi-document summarization [16]. Controllable generation methods, like contrastive prefix fine-tuning [17], and personalized content generation via text-to-image diffusion [6] and video generation [5], enhance precision. Multilingual transformers [18] improve LLM understanding for accurate reports.

Integrating visual information is crucial for medical image report generation, building on image captioning [19]. Vision-guided generative LMs [20] and VLMs, using visual in-context learning [21], are critical for interpreting medical imagery. Efficient integration includes vision representation compression [22]. Model training optimizations, like loss distillation [23], contribute to robust models. LLMs identify visual entities [24], while VLMs are enhanced for ambiguity resolution [25] and visual reasoning [26]. The field advances towards multimodal agent intelligence for medical diagnosis, using role-specialized collaboration [27] and broader multimodal agent intelligence [4].

Responsible deployment of LLMs and VLMs requires effective interaction and ethics. Prompt engineering [28] is vital for controlling generative models. Ensuring model robustness and interpretability, e.g., in face anti-spoofing [1], is paramount for fairness and accuracy in medical report generation. Addressing bias, like with Auto-Debias [29], is also crucial for fairness and accuracy.

2.2. Graph Neural Networks and Temporal Modeling in Medical Imaging

Medical imaging analysis benefits from GNNs and temporal modeling, capturing structural relationships and dynamic changes.

GNNs are powerful for non-Euclidean data; GCNs capture complex inter-dependencies, e.g., in information extraction [30]. GATs enhance interpretability and performance via differential node weighting [31], applicable to medical image analysis. Cross-modal attention frameworks [32] integrate diverse multimodal medical data.

Longitudinal medical imaging requires robust temporal modeling. Hierarchical attentional hybrid networks [33] combine CNNs, GRUs, and attention for dynamic data. Integrating historical context, e.g., AR-RAG [34], is crucial for predictions. Temporal misalignment [35] and adaptation [36] are critical challenges in time-series data, impacting medical image analysis like lesion tracking.

Integrating GNNs with temporal modeling forms spatio-temporal graph approaches, addressing spatial dependencies and temporal dynamics. The MTAG model [37] captures multimodal and temporal interactions, foundational for spatio-temporal graph representations in medical imaging for tasks like disease progression prediction. This convergence promises advanced medical imaging analysis.

3. Method

Our proposed **TempSem-GraphNet** is an end-to-end framework meticulously designed for generating coherent and accurate radiology reports from multi-temporal Chest X-ray (CXR) images. The core idea behind this innovative approach is to explicitly model the dynamic evolution of pathologies and their associated clinical context. This is achieved by constructing a novel **Multi-modal Temporal-Semantic Graph** (TempSem-Graph), which comprehensively unifies diverse visual findings extracted from serial CXRs with rich semantic concepts derived from corresponding historical radiology reports. This multifaceted graph representation, encapsulating intricate temporal and semantic relationships, is subsequently processed by a specialized graph attention network. This network is tasked with extracting profound contextual information, which then serves as a guiding mechanism for a carefully fine-tuned Large Language Model (LLM) to synthesize the final, comprehensive radiology report.

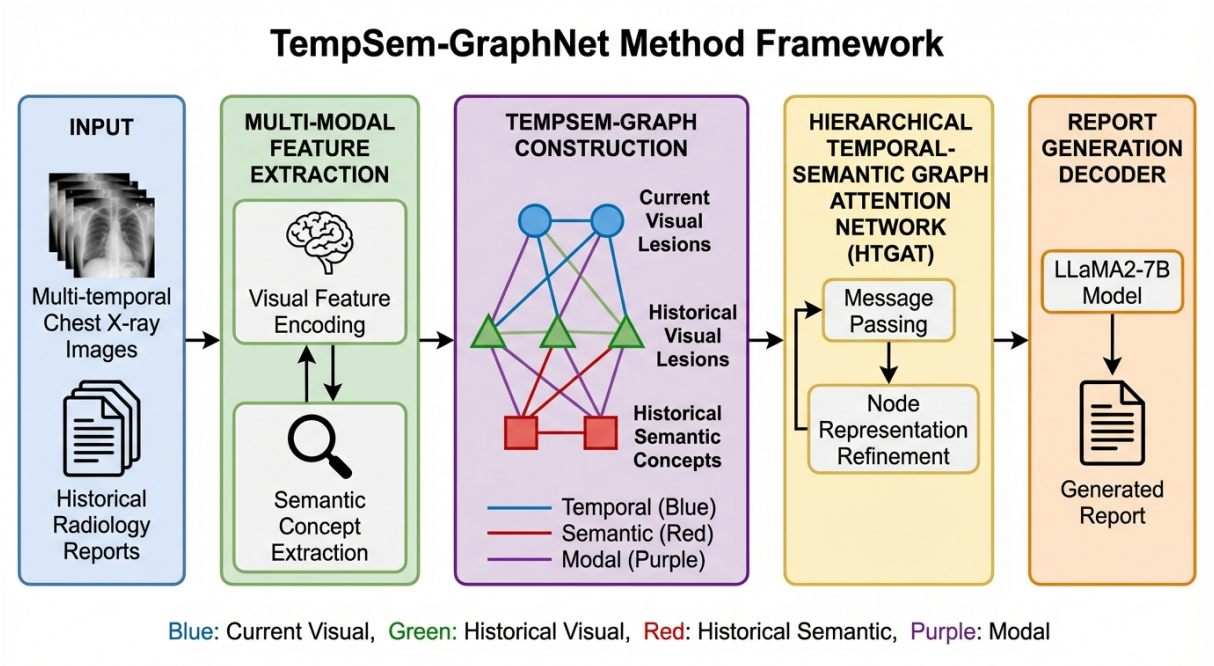


Figure 2: Overview of the **TempSem-GraphNet** framework. The pipeline illustrates the multi-modal input processing, TempSem-Graph construction from extracted visual lesions and semantic concepts, contextual information aggregation via a Hierarchical Temporal-Semantic Graph Attention Network (HTGAT), and final radiology report generation by a fine-tuned LLaMA2-7B decoder.

3.1. Overall Framework

The **TempSem-GraphNet** framework takes as input a series of CXR images, which typically comprises a current diagnostic image alongside one or more preceding historical images, along with their corresponding historical radiology reports. The framework's operational pipeline is structured into three principal stages. First, a dedicated multi-modal feature extraction module processes both visual inputs (CXR images) and textual inputs (historical reports) to robustly identify salient visual lesions and pertinent semantic medical concepts. Second, these meticulously extracted multi-modal entities serve as the foundational elements for constructing the **TempSem-Graph**. Within this graph, nodes distinctly represent either visual lesions or semantic concepts, while edges are precisely engineered to encode complex temporal, semantic, and inter-modal relationships between these entities. Third, a **Hierarchical Temporal-Semantic Graph Attention Network** (HTGAT) performs multi-layer message passing on this intricately structured graph. This process is designed to adaptively aggregate contextual information from both local neighborhoods and broader graph structures, culminating in a comprehensive and context-rich graph embedding. Finally, this aggregated graph embedding is seamlessly integrated into a fine-tuned LLaMA2-7B model, which functions as a powerful and highly specialized decoder responsible for generating the natural language radiology report.

3.2. Multi-modal Feature Extraction

To obtain a thorough and comprehensive understanding of the patient's longitudinal medical condition, our framework performs robust feature extraction from both visual (imaging) and textual (report) modalities. This dual-modality approach ensures that both observed morphological changes and documented clinical narratives are leveraged.

3.2.1. Visual Feature Encoding and Lesion Detection For the encoding of visual features from CXR images, we employ a frozen 2D pre-trained Vision Transformer (ViT) as our foundational visual backbone. Given a current CXR image I_C and a set of historical CXR images $\{I_{H,t}\}_{t=1}^{T-1}$ representing different time points, the ViT systematically extracts two primary types of features for each image: a global visual feature $f_G \in \mathbb{R}^{D_G}$ capturing overall image context, and a set of localized patch-level features $f_P \in \mathbb{R}^{N_P \times D_P}$ providing granular regional information.

To precisely identify and localize specific pathological regions, we integrate a pre-trained multi-modal lesion detection model, such as one based on YOLOv7 or RetinaNet. This detector processes each individual CXR image $I \in \{I_C, I_{H,t}\}$ to output a collection of bounding boxes $\mathcal{B} = \{b_j\}_{j=1}^{K}$. Each bounding box b_j is characterized by a tuple $(x_j, y_j, w_j, h_j, c_j, s_j)$, which specifies its top-left coordinates (x_j, y_j), width w_j, height w_j, the predicted lesion category c_j, and its associated confidence score s_j. For each detected lesion j, a dedicated visual feature $v_j \in \mathbb{R}^{D_V}$ is extracted. This is achieved by applying a pooling mechanism, such as ROI-Align, to aggregate the ViT patch features situated within its defined bounding box. This pooled feature is then concatenated with the normalized bounding box coordinates and an embedding corresponding to its predicted category, forming a comprehensive visual representation for each lesion.

3.2.2. Semantic Concept Extraction From the provided historical radiology reports $\{R_{H,t}\}_{t=1}^{T-1}$, we systematically extract key medical entities and relevant clinical concepts. This critical step is accomplished using advanced natural language processing (NLP) techniques, which may include sophisticated named entity recognition (NER) models specifically trained on medical corpora, or highly optimized rule-based keyword extraction algorithms. Each identified concept s_k is assigned a unique identifier and is subsequently represented by its corresponding dense word embedding $e_{s_k} \in \mathbb{R}^{D_S}$. These embeddings are typically obtained from a pre-trained medical text embedding model, such as BioClinicalBERT, ensuring domain-specific semantic representation. This comprehensive process yields an exhaustive set of pertinent semantic concepts $\mathcal{S}_H = \{s_k\}_{k=1}^{M}$ aggregated from all available historical reports.

3.3. Multi-modal Temporal-Semantic Graph (TempSem-Graph) Construction

The very core of our proposed approach resides in the meticulous construction of the **TempSem-Graph** $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. This graph is designed to explicitly and comprehensively model the intricate relationships and dynamic interactions between visual lesions identified in imaging data and semantic concepts derived from textual reports, all meticulously organized across different time points. This heterogeneous graph structure allows for a holistic understanding of patient history and pathology evolution.

3.3.1. Nodes $\mathcal{N}$ The nodes forming the graph $\mathcal{N}$ represent distinct individual entities meticulously extracted from the multi-modal inputs. These nodes are categorized into three primary types:

Current Visual Lesion Nodes ($\mathcal{V}_C$): Each node denoted as $n_{c,i} \in \mathcal{V}_C$ specifically corresponds to a unique lesion that has been detected in the current CXR image I_C. Its initial feature vector, $x_{c,i}$, is directly derived from the comprehensive visual feature v_i as described in Section 2.2.1.

Historical Visual Lesion Nodes ($\mathcal{V}_H$): Similarly, each node $n_{h,j} \in \mathcal{V}_H$ corresponds to a distinct lesion that was identified in one of the historical CXR images $I_{H,t}$. Its initial feature vector, $x_{h,j}$, is analogously derived from the historical image's visual features, ensuring consistency in representation.

Historical Semantic Concept Nodes ($\mathcal{S}_H$): Each node $n_{s,k} \in \mathcal{S}_H$ embodies a key medical concept or entity that was extracted from a historical radiology report $R_{H,t}$. Its initial feature

vector, $x_{s,k}$, is the rich semantic embedding e_{s_k} as detailed in Section 2.2.2.

The complete and unified set of all nodes within the **TempSem-Graph** is thus defined as $\mathcal{N} = \mathcal{V}_C \cup \mathcal{V}_H \cup \mathcal{S}_H$.

3.3.2. Edges $\mathcal{E}$ Edges within the **TempSem-Graph** are judiciously established between nodes to precisely capture and encode different types of relationships, thereby forming a rich relational structure:

Temporal Edges ($\mathcal{E}_T$): These edges are crucial for modeling the longitudinal evolution of pathologies. They specifically connect visual lesion nodes across different time points, thereby indicating the temporal progression, regression, or stability of a particular pathological finding. An edge $(n_{c,i}, n_{h,j}) \in \mathcal{E}_T$ is instantiated if the current lesion $n_{c,i} \in \mathcal{V}_C$ and the historical lesion $n_{h,j} \in \mathcal{V}_H$ are determined to represent the same underlying pathology or an evolving manifestation of it. This determination is made based on a set of criteria including, but not limited to, significant spatial overlap (quantified, for instance, by Intersection over Union, IoU) between their respective bounding boxes, a high degree of similarity in their predicted lesion categories, and their temporal proximity.

Semantic Edges ($\mathcal{E}_S$): These edges capture semantic associations within the graph. They connect semantic concept nodes to each other if they exhibit a meaningful semantic relationship (e.g., 'pneumonia' and 'infiltrate' signifying a related clinical condition or manifestation). Furthermore, semantic edges also connect visual lesion nodes to semantic concept nodes when the concept directly describes or is strongly associated with the visual finding. For example, an edge $(n_{h,j}, n_{s,k}) \in \mathcal{E}_S$ is established if the lesion $n_{h,j}$ was explicitly described or mentioned by the concept $n_{s,k}$ in its original historical report, or if $n_{s,k}$ is a widely recognized medical descriptor for the specific category of the lesion $n_{h,j}$.

Modal Edges ($\mathcal{E}_M$): These edges are designed to bridge the gap between different modalities, linking visual lesions with their corresponding textual descriptions or related clinical narratives, even if these appear in distinct input sources or across time. Specifically, they connect current visual lesion nodes $n_{c,i}$ (originating from the current CXR image) to relevant historical semantic concept nodes $n_{s,k}$ (extracted from historical reports). This connection is formed if an initial description of that lesion, or a similar underlying pathology, is identified within the patient's historical reports, thereby facilitating a comprehensive multi-modal context for interpreting current imaging findings.

Each edge $e \in \mathcal{E}$ is also enriched by being augmented with a distinct type embedding or a dynamically learned weight, which reflects the specific nature and strength of the relationship it represents. The initial features of all nodes x_i are collectively organized into a feature matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{N}| \times D}$, and the intricate connectivity patterns are comprehensively represented by an adjacency matrix $\mathbf{A}$.

3.4. Hierarchical Temporal-Semantic Graph Attention Network (HTGAT)

The intricately constructed **TempSem-Graph** $\mathcal{G}$ is subsequently processed by a **Hierarchical Temporal-Semantic Graph Attention Network** (HTGAT). The primary objective of HTGAT is to judiciously aggregate information across the graph and learn highly context-aware, refined node representations. HTGAT is specifically engineered to perform multi-layer message passing, thereby facilitating the effective flow of information across diverse node types and heterogeneous edge types. This advanced design enables the network to meticulously capture complex interactions, explicit temporal dependencies, and nuanced semantic relationships embedded within the graph structure.

In each successive layer l of the HTGAT, the feature representation of a given node n_i is adaptively updated by intelligently aggregating information from its immediate neighbors $\mathcal{N}(n_i)$. The inherent attention mechanism is a critical component, allowing the network to assign varying

degrees of importance or relevance to different neighboring nodes and distinct edge types during the aggregation process. Specifically, for a node n_i possessing a feature vector $\mathbf{h}_i^{(l)}$ at layer l, its new, updated feature vector $\mathbf{h}_i^{(l+1)}$ is computed through the following sequence of operations:

$$\mathbf{z}_i^{(l)} = \sum_{n_j \in \mathcal{N}(n_i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \tag{1}$$

$$\mathbf{h}_i^{(l+1)} = \sigma(\mathbf{z}_i^{(l)}) \tag{2}$$

In these equations, $\mathbf{W}^{(l)}$ represents a learnable weight matrix specific to layer l, which projects the neighbor's features into a common space. The function σ denotes an element-wise activation function, such as LeakyReLU, applied to introduce non-linearity. Crucially, $\alpha_{ij}^{(l)}$ signifies the attention coefficient, a scalar weight calculated for the directed edge connecting node n_i to its neighbor n_j. These attention coefficients are dynamically computed based on the current feature representations of node n_i, node n_j, and the specific type of edge e_{ij} connecting them:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^{(l)T}\left[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(l)}\mathbf{h}_j^{(l)} \| \mathbf{e}_{e_{ij}}\right]\right)\right)}{\sum_{n_k \in \mathcal{N}(n_i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^{(l)T}\left[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(l)}\mathbf{h}_k^{(l)} \| \mathbf{e}_{e_{ik}}\right]\right)\right)} \tag{3}$$

Here, $\mathbf{a}^{(l)}$ is a learnable attention vector specific to layer l, which is used to compute the attention score. The operator $\|$ denotes the concatenation operation, combining the transformed features of the central node, its neighbor, and the edge type embedding. The term $\mathbf{e}_{e_{ij}}$ represents a learnable embedding that uniquely encodes the specific type of the edge e_{ij} (i.e., whether it is a temporal, semantic, or modal relationship). The HTGAT framework further enhances its representational power by employing multiple attention heads, allowing it to capture diverse relational patterns simultaneously, and by utilizing a hierarchical structure to process information at varying granularities and effectively fuse heterogeneous signals. After L layers of message passing, the network yields refined node embeddings $\mathbf{h}_i^{(L)}$ for all nodes in the graph. These final embeddings comprehensively capture the intricate visual, temporal, and semantic context pertinent to each entity. A global graph representation, denoted as $\mathbf{h}_\mathcal{G}$, is subsequently derived by applying a suitable pooling mechanism (e.g., mean pooling or max pooling) over all these final node embeddings, effectively summarizing the entire graph's learned context.

3.5. Report Generation Decoder

The global graph representation $\mathbf{h}_\mathcal{G}$, which encapsulates the distilled multi-modal, temporal, and semantic context, is subsequently channeled into a Large Language Model (LLM) for the ultimate task of generating the radiology report. We judiciously utilize the LLaMA2-7B model as our robust decoder, having fine-tuned it extensively to ensure the generation of clinically coherent, accurate, and fluent reports that align with medical standards.

To effectively integrate the rich contextual information from the **TempSem-Graph** into the LLM, the aggregated graph feature $\mathbf{h}_\mathcal{G}$ is first projected into a sequence of continuous "soft prompts" $\mathbf{P} \in \mathbb{R}^{S \times D_{LLM}}$. This sequence of soft prompts is then strategically prepended to the input token embeddings of the LLM. This mechanism allows the LLM to condition its generative process directly on the extracted contextual information, thereby steering its output towards producing reports that are highly relevant to the patient's imaging findings and historical data. Following this conditioning, the LLM proceeds to generate the radiology report token by token, leveraging its vast linguistic knowledge base, conditioned by both the soft prompts and the sequence of previously generated tokens.

For the purpose of efficient and effective fine-tuning of the LLaMA2-7B model on our specialized medical task, we employ the Low-Rank Adaptation (LoRA) technique. Specifically, LoRA matrices are applied to the linear query and value projection layers within the LLM's multi-head attention mechanism, which are critical for capturing input relationships. Furthermore, to allow for greater task-specific adaptation, the embedding layer responsible for initial token representations and the final output head of the LLM are configured to be trainable parameters. Our particular LoRA configuration utilizes a rank of 64, an alpha scaling factor of 64, and incorporates a dropout rate of 0.1 to prevent overfitting. The entire model is optimized using the AdamW optimizer with a learning rate of 3e-5, minimizing the negative log-likelihood of the target radiology reports. This carefully designed fine-tuning strategy enables the LLM to adapt powerfully and effectively to the domain-specific nuances of CXR report generation, significantly mitigating the computational costs typically associated with full model re-training.

4. Experiments

In this section, we present the experimental setup, evaluate the performance of our proposed **TempSem-GraphNet** framework, and conduct extensive ablation studies to validate the effectiveness of its key components. We also include a human evaluation to assess the clinical utility and coherence of the generated reports from a radiologist's perspective.

4.1. Datasets

We conducted our experiments on two widely utilized datasets in medical imaging:

- **MIMIC-CXR-JPG** [13]: This is a large-scale, publicly available dataset comprising 377,110 chest X-ray images associated with 227,835 imaging studies and 206,561 radiology reports from 65,379 distinct patients. Critically, it contains numerous serial CXR examinations for individual patients, making it ideal for evaluating temporal modeling capabilities. We follow standard data splits, using 70% for training, 10% for validation, and 20% for testing.
- **CheXpert** [14]: Consisting of 224,316 chest radiographs of 65,240 patients, each labeled for the presence of 14 observations, this dataset was primarily used for pre-training our visual encoder and developing the multi-modal lesion detection model.

4.2. Evaluation Metrics

To provide a comprehensive evaluation of the generated radiology reports, we employ two sets of metrics:

- **Natural Language Generation (NLG) Metrics**: These metrics assess the fluency, grammatical correctness, and content overlap of the generated reports with expert-written reference reports. We report standard metrics including BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR. Higher scores indicate better performance.
- **Clinical Entity (CE) Metrics**: To evaluate the clinical relevance, factual correctness, and completeness of medical findings in the generated reports, we extract clinical entities (e.g., diseases, anatomical locations, findings) using a pre-trained clinical NLP parser and compare them against those extracted from reference reports. We report Precision (P), Recall (R), and F1-score (F1) for these entities. Higher F1 scores denote more accurate and comprehensive clinical information.

4.3. Implementation Details

All CXR images were resized to 512×512 pixels and normalized according to the pre-trained visual encoder's requirements. Our visual feature encoder, a frozen 2D pre-trained Vision

Transformer (ViT), was initialized from weights pre-trained on ImageNet and further fine-tuned on the CheXpert dataset for medical image understanding. The multi-modal lesion detection and semantic extractor utilized a YOLOv7-based model, pre-trained on a subset of CheXpert and MIMIC-CXR-JPG annotations. For semantic concept extraction, we employed a medical Named Entity Recognition (NER) model built on BioClinicalBERT.

The **TempSem-Graph** was constructed by defining temporal edges for lesions with IoU ≥ 0.5 and similar categories across time points, semantic edges based on cosine similarity ≥ 0.7 between concept embeddings or direct mentions in reports, and modal edges by linking current visual lesions to historical semantic concepts describing similar pathologies. Our Hierarchical Temporal-Semantic Graph Attention Network (HTGAT) consisted of 2 graph attention layers.

For the report generation decoder, we used **LLaMA2-7B** and fine-tuned it efficiently using LoRA [12]. The LoRA configuration involved a rank of 64, an alpha scaling factor of 64, and a dropout rate of 0.1. Crucially, the embedding layer and the output head of the LLaMA2-7B model were set to be trainable parameters to enhance task-specific adaptation. The model was optimized using the AdamW optimizer with a learning rate of 3e-5 for 8 epochs. This training regimen took approximately 15 hours on a single NVIDIA A100 GPU. The auxiliary task of multi-modal lesion tracking and description was trained for 50 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-5.

4.4. Baselines

We compare our proposed **TempSem-GraphNet** with several strong baseline methods to demonstrate its superior performance:

- **Image-to-Sequence (Baseline)**: This represents a conventional approach where a visual encoder extracts features from a single current CXR image, which are then fed into a sequence decoder (e.g., an LSTM or Transformer) to generate the report. This method does not incorporate historical images or explicit graph structures.
- **Visual-GraphNet (No Temporal)**: An extension of graph-based methods, this baseline constructs a graph solely from visual lesions detected in the **current** CXR image. It uses a graph neural network to aggregate spatial relationships but lacks any temporal or semantic information from historical data. This baseline assesses the impact of explicit temporal and multi-modal semantic modeling.
- **LLM-Prompting (Features Only)**: This baseline utilizes the same frozen visual encoder and lesion detector as our method, but instead of constructing a graph, it directly concatenates global image features and lesion features into a single vector, which is then projected into soft prompts for the LLaMA2-7B LLM. This method evaluates the advantage of structured graph representation over simple feature concatenation for LLM conditioning.

4.5. Main Results

Table 1 presents the quantitative comparison of our proposed **TempSem-GraphNet** with the baseline methods on the MIMIC-CXR-JPG test set.

As shown in Table 1, our proposed **TempSem-GraphNet** significantly outperforms all baseline methods across all Natural Language Generation (NLG) and Clinical Entity (CE) metrics. Specifically, TempSem-GraphNet achieves a BLEU-4 score of **0.2441** and an F1-score of **0.288** for clinical entity extraction, demonstrating superior performance in both linguistic quality and clinical accuracy.

Compared to the **Image-to-Sequence (Baseline)**, which only considers the current image, our method shows substantial improvements, highlighting the critical role of integrating multi-temporal information. The **Visual-GraphNet (No Temporal)**, while benefiting from a graph structure for current visual lesions, still falls short of our method. This performance gap

Table 1: Performance comparison of **TempSem-GraphNet** with existing methods on CXR report generation. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Higher scores are better for all metrics. Bold indicates the best performance.

Method	B-1	B-2	B-3	B-4	R-L	METEOR	P (CE)	R (CE)	F1 (CE)
Image-to-Sequence (Baseline)	0.4552	0.3489	0.2761	0.2230	0.2985	0.3957	0.281	0.155	0.198
Visual-GraphNet (No Temporal)	0.4687	0.3601	0.2872	0.2335	0.3051	0.4072	0.312	0.188	0.235
LLM-Prompting (Features Only)	0.4735	0.3648	0.2910	0.2378	0.3090	0.4120	0.335	0.199	0.248
TempSem-GraphNet (Ours)	**0.4812**	**0.3725**	**0.2988**	**0.2441**	**0.3155**	**0.4187**	**0.368**	**0.235**	**0.288**

Table 2: Ablation study on key components of **TempSem-GraphNet**. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Bold indicates the best performance for the full model.

Method	B-1	B-2	B-3	B-4	R-L	METEOR	P (CE)	R (CE)	F1 (CE)
TempSem-GraphNet (Full)	**0.4812**	**0.3725**	**0.2988**	**0.2441**	**0.3155**	**0.4187**	**0.368**	**0.235**	**0.288**
w/o Temporal Edges ($\mathcal{E}_T$)	0.4678	0.3591	0.2865	0.2330	0.3045	0.4068	0.321	0.198	0.245
w/o Semantic Edges ($\mathcal{E}_S$)	0.4705	0.3620	0.2888	0.2355	0.3072	0.4095	0.339	0.205	0.255
w/o Modal Edges ($\mathcal{E}_M$)	0.4752	0.3670	0.2935	0.2400	0.3115	0.4135	0.348	0.218	0.267
w/o HTGAT (Simple Pool)	0.4610	0.3540	0.2810	0.2285	0.3005	0.4010	0.305	0.180	0.226
w/o LoRA Fine-tuning	0.4490	0.3425	0.2700	0.2180	0.2930	0.3890	0.270	0.145	0.187

underscores the importance of explicitly modeling temporal relationships between lesions across different time points and incorporating semantic context from historical reports. Furthermore, the **LLM-Prompting (Features Only)** baseline, despite utilizing the powerful LLaMA2-7B, performs worse than TempSem-GraphNet. This indicates that simply concatenating features into prompts is less effective than our structured **TempSem-Graph** and HTGAT for aggregating complex multi-modal, temporal, and semantic information, proving that the explicit graph representation provides a more coherent and comprehensive context for the LLM. These results collectively validate the efficacy of our novel multi-modal temporal-semantic graph construction and hierarchical graph attention network in generating more coherent, accurate, and clinically informative CXR reports.

4.6. Ablation Study

To thoroughly understand the contribution of each component within our **TempSem-GraphNet** framework, we conducted a comprehensive ablation study. The results are summarized in Table 2.

Effect of Temporal Edges ($\mathcal{E}_T$): When temporal edges are removed from the TempSem-Graph, the model's performance drops noticeably across all metrics, with a decrease of 0.0111 in BLEU-4 and 0.043 in F1 (CE). This highlights the critical role of explicitly modeling lesion evolution over time for generating coherent reports, especially for describing dynamic changes.

Effect of Semantic Edges ($\mathcal{E}_S$): Removing semantic edges, which link related concepts and visual findings to their descriptions, leads to a reduction of 0.0086 in BLEU-4 and 0.033 in F1 (CE). This indicates that the explicit semantic relationships are vital for enhancing the factual accuracy and clinical completeness of the generated reports.

Effect of Modal Edges ($\mathcal{E}_M$): The absence of modal edges, bridging current visual lesions with historical semantic concepts, results in a decrease of 0.0041 in BLEU-4 and 0.021 in F1 (CE). While less severe than temporal or semantic edges, this still signifies the importance of cross-modal linking for providing a comprehensive historical context to current findings.

Effect of HTGAT (Simple Pool): Replacing the Hierarchical Temporal-Semantic Graph

Table 3: Average scores from human evaluation by three radiologists (1-5 Likert scale). Higher scores indicate better quality. Bold indicates the best performance.

Method	Clinical Accuracy	Temporal Coherence	Report Fluency	Clinical Utility
LLM-Prompting (Features Only)	3.4	2.9	3.8	3.1
TempSem-GraphNet (Ours)	**4.2**	**4.1**	**4.3**	**4.1**

Attention Network (HTGAT) with a simple mean pooling mechanism over node features after graph construction (labeled "w/o HTGAT") leads to a substantial performance drop across all metrics (e.g., 0.0156 decrease in BLEU-4 and 0.062 decrease in F1 (CE)). This unequivocally demonstrates the effectiveness of HTGAT's multi-layer message passing and attention mechanisms in aggregating context-aware, refined node representations.

Effect of LoRA Fine-tuning: To evaluate the contribution of efficient LLM fine-tuning, we conducted an experiment without LoRA, which essentially means the LLM decoder received features but was not adaptively fine-tuned on the task. This leads to a severe degradation in performance across all metrics (e.g., 0.0261 decrease in BLEU-4 and 0.101 decrease in F1 (CE)). This underscores the necessity of task-specific adaptation for the LLM decoder to effectively translate aggregated graph features into high-quality medical reports.

These ablation results confirm that each component of **TempSem-GraphNet**, particularly the different types of edges in the TempSem-Graph and the HTGAT for graph processing, significantly contributes to the model's overall superior performance in generating coherent and clinically accurate CXR reports.

4.7. Human Evaluation

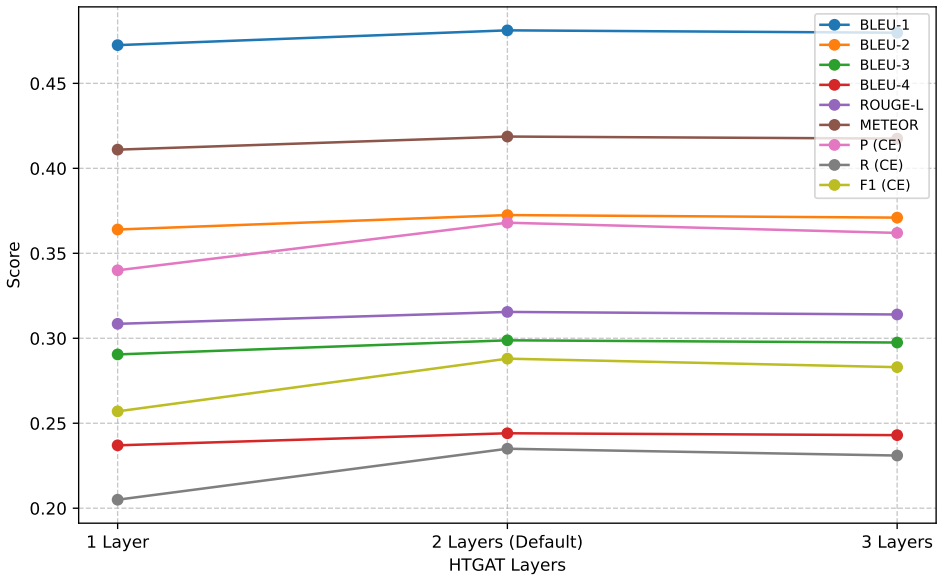
To further assess the clinical utility and qualitative aspects of the generated reports, we conducted a human evaluation involving three board-certified radiologists. They were presented with randomly selected CXR image sets (current and historical) and corresponding reports generated by **TempSem-GraphNet** and the best-performing baseline, **LLM-Prompting (Features Only)**, without knowing which model generated which report. Radiologists rated each report on a Likert scale from 1 (poor) to 5 (excellent) across four key criteria:

- **Clinical Accuracy**: How well the report accurately describes the observed findings and their evolution.
- **Temporal Coherence**: How consistently and logically the report describes changes over time, referencing historical context appropriately.
- **Report Fluency**: The naturalness, readability, and grammatical correctness of the language.
- **Clinical Utility**: Overall usefulness of the report for clinical decision-making.

Table 3 presents the average scores from the radiologists.

The human evaluation results corroborate our quantitative findings. **TempSem-GraphNet** consistently received significantly higher ratings across all qualitative criteria compared to the **LLM-Prompting (Features Only)** baseline. Notably, our method showed a substantial improvement in "Temporal Coherence" (4.1 vs. 2.9), directly reflecting its ability to model and describe the evolution of pathologies over time. Furthermore, "Clinical Accuracy" and "Clinical Utility" scores were markedly higher for TempSem-GraphNet, indicating that the reports generated by our framework are not only linguistically sound but also more factually correct and clinically valuable to radiologists. These qualitative assessments provide strong evidence that our explicit modeling of multi-modal temporal-semantic relationships translates into genuinely more coherent and clinically actionable radiology reports.

Figure 3: Performance of **TempSem-GraphNet** with varying HTGAT depths. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Bold indicates the best performance among different depths.



4.8. Impact of HTGAT Depth

The depth of the Graph Attention Network plays a crucial role in its ability to aggregate information from distant nodes within the graph. To investigate the optimal number of layers for our Hierarchical Temporal-Semantic Graph Attention Network (HTGAT), we evaluated the performance of **TempSem-GraphNet** using 1, 2 (our default), and 3 graph attention layers. Increasing the number of layers allows information to propagate further, potentially capturing more global context, but can also lead to issues like over-smoothing, where node representations become indistinguishable.

Figure 3 presents the performance metrics for different HTGAT depths.

The results indicate that 2 layers yield the optimal performance across most metrics. A single layer of HTGAT, while offering an improvement over simple pooling (as seen in the ablation study), does not capture sufficient long-range dependencies within the TempSem-Graph. Increasing the depth to 3 layers shows a slight decrease in performance compared to 2 layers. This suggests that for the complexity of our TempSem-Graph and the density of relationships, 2 layers strike a good balance, effectively aggregating contextual information without suffering from over-smoothing or introducing excessive noise from overly distant graph regions.

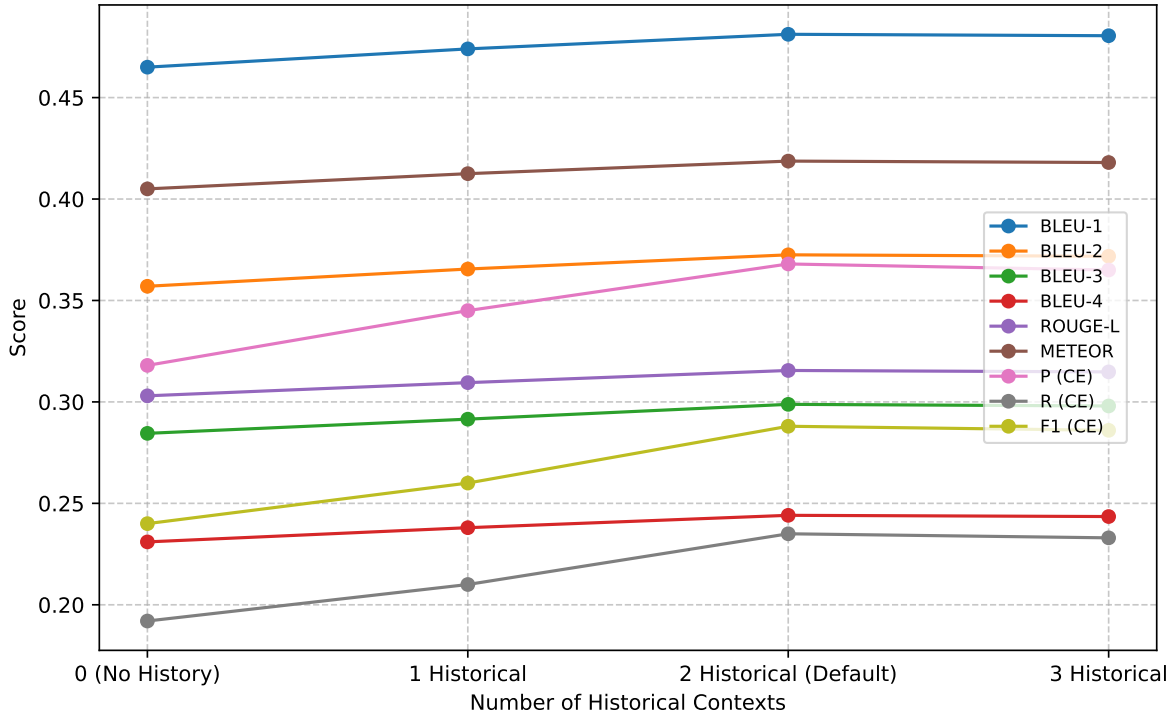
4.9. Sensitivity to Historical Context Availability

A cornerstone of **TempSem-GraphNet** is its ability to leverage multi-temporal information. To understand how the quantity of historical context influences report generation quality, we evaluated our model under various conditions of historical data availability. This included scenarios with no historical data, a single historical CXR and report, our default setup of two historical CXRs and reports, and up to three historical CXRs and reports.

Figure 4 summarizes the performance across these different levels of historical context. "Hist. Data" refers to the number of preceding CXR images and their corresponding reports used as input.

The results clearly demonstrate that incorporating historical context significantly enhances

Figure 4: Sensitivity analysis of **TempSem-GraphNet** to the volume of historical context. Hist. Data: Number of historical CXR images and reports used. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Bold indicates the best performance among different historical data volumes.



performance across all metrics. Using even a single historical examination provides a substantial boost over having no history, particularly in clinical entity metrics, signifying improved factual accuracy due to context. Performance generally improves as more historical data is introduced, peaking at 2 historical examinations. Using 3 historical examinations shows a marginal decrease, which could be attributed to increased graph complexity and potential noise, or the diminishing returns of very distant historical data given the typical progression patterns of acute chest pathologies. This analysis validates our framework's capacity to effectively leverage multi-temporal inputs, with an optimal balance achieved around two preceding studies for generating the most comprehensive and accurate reports.

4.10. Analysis of Graph Construction Thresholds

The accurate construction of the **TempSem-Graph** heavily relies on specific thresholds used to establish relationships between nodes. To evaluate the robustness and sensitivity of our model to these parameters, we conducted experiments varying the Intersection over Union (IoU) threshold for temporal edges and the cosine similarity threshold for semantic edges.

4.10.1. Temporal Edge IoU Threshold Temporal edges connect visual lesions across different time points. A higher IoU threshold implies a stricter criterion for considering two lesions as the "same" evolving entity, potentially missing subtle changes, while a lower threshold might mistakenly link unrelated findings. Our default IoU threshold for temporal edges ($\mathcal{E}_T$) is 0.5. We tested thresholds of 0.3, 0.5, and 0.7.

Table 4: Performance of **TempSem-GraphNet** with varying Temporal Edge IoU Thresholds. IoU Thresh.: Intersection over Union Threshold. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Bold indicates the best performance among different thresholds.

IoU Thresh.	B-1	B-2	B-3	B-4	R-L	METEOR	P (CE)	R (CE)	F1 (CE)
0.3	0.4760	0.3675	0.2938	0.2405	0.3120	0.4140	0.350	0.220	0.269
0.5 (Default)	**0.4812**	**0.3725**	**0.2988**	**0.2441**	**0.3155**	**0.4187**	**0.368**	**0.235**	**0.288**
0.7	0.4785	0.3700	0.2965	0.2420	0.3135	0.4165	0.358	0.228	0.280

Table 5: Performance of **TempSem-GraphNet** with varying Semantic Edge Cosine Similarity Thresholds. Cos Sim Thresh.: Cosine Similarity Threshold. P (CE): Clinical Entity Precision, R (CE): Clinical Entity Recall, F1 (CE): Clinical Entity F1-score. Bold indicates the best performance among different thresholds.

Cos Sim Thresh.	B-1	B-2	B-3	B-4	R-L	METEOR	P (CE)	R (CE)	F1 (CE)
0.5	0.4770	0.3685	0.2945	0.2410	0.3125	0.4148	0.355	0.222	0.272
0.7 (Default)	**0.4812**	**0.3725**	**0.2988**	**0.2441**	**0.3155**	**0.4187**	**0.368**	**0.235**	**0.288**
0.9	0.4795	0.3705	0.2970	0.2425	0.3140	0.4170	0.360	0.230	0.282

Table 4 shows the impact of varying the IoU threshold for temporal edge creation.

An IoU threshold of 0.5 demonstrates the optimal balance, yielding the highest performance across all metrics. A lower threshold (0.3) likely introduces spurious temporal connections between unrelated lesions, leading to decreased accuracy. Conversely, a higher threshold (0.7) might be too restrictive, failing to link truly evolving pathologies due to slight shifts in bounding box detections, thereby reducing the temporal coherence captured.

4.10.2. Semantic Edge Cosine Similarity Threshold Semantic edges connect related semantic concepts or link visual lesions to their textual descriptions. The cosine similarity threshold dictates how semantically similar two entities must be to form an edge. A lower threshold could introduce noisy or irrelevant connections, while a higher threshold might prevent valuable associations from being formed. Our default cosine similarity threshold for semantic edges ($\mathcal{E}_S$) is 0.7. We experimented with thresholds of 0.5, 0.7, and 0.9.

Table 5 shows the impact of varying the cosine similarity threshold for semantic edge creation.

Similar to temporal edges, a cosine similarity threshold of 0.7 for semantic edges achieves the best results. A lower threshold (0.5) increases the density of semantic edges, but likely includes weak or misleading associations, causing a slight performance drop. Conversely, a higher threshold (0.9) is too stringent, potentially severing valid semantic connections and reducing the richness of the graph's semantic context, also leading to a slight performance decrease. These experiments confirm the careful calibration of our graph construction parameters is vital for the optimal functioning of **TempSem-GraphNet**.

4.11. Qualitative Analysis and Error Modes

Beyond quantitative metrics and human evaluation, we conducted an in-depth qualitative analysis of generated reports from **TempSem-GraphNet** and the best baseline (**LLM-Prompting (Features Only)**) to gain insights into specific strengths and identify recurring error patterns. This involved reviewing a random subset of 100 generated reports from each method against the reference reports and corresponding images.

Table 6: Key qualitative strengths observed in **TempSem-GraphNet** generated reports compared to baselines.

Strength	Description
Precise Temporal Sequencing	Accurately describes the progression, regression, or stability of findings across different time points (e.g., "right lower lobe infiltrate has slightly improved since prior study").
Contextualized Finding Descriptions	Integrates historical context into the description of current findings, providing a richer clinical picture (e.g., "mild cardiomegaly, stable from prior" rather than just "mild cardiomegaly").
Reduced Redundancy	Less prone to repeating information across sentences or sections, as the graph structure helps summarize and prioritize unique findings and changes.
Factual Consistency	Higher adherence to factual information present in the input images and historical reports, particularly for complex or evolving pathologies.
Coherent Negation	More consistent and accurate use of negations when findings are absent or resolved, driven by clearer tracking of pathologies over time.

Table 7: Common error modes observed in reports generated by **TempSem-GraphNet**.

Error Mode	Description
Subtle Finding Omission	Occasionally misses very subtle or ambiguous findings that might be inferred by a human radiologist.
Minor Hallucinations	Infrequently generates minor findings or details not present in the images or historical reports, though less frequent than baselines.
Overgeneralization	Sometimes describes findings in a slightly more general way than the specific nuances described in a human-written report.
Imprecise Measurements	Struggles with quantifying exact measurements or degrees of change (e.g., "mild" vs. "moderate," or precise size changes), often relying on qualitative terms.
Complex Case Simplification	For extremely complex multi-pathology cases, the report may simplify the intricate interactions or causal relationships between findings.

Table 6 summarizes the key qualitative strengths observed in reports generated by **TempSem-GraphNet**, particularly in comparison to baseline methods.

Table 7 outlines common error modes observed in the reports generated by **TempSem-GraphNet**. While performing significantly better, our model is not immune to challenges.

This qualitative analysis confirms that the explicit modeling of temporal and semantic relationships in **TempSem-GraphNet** significantly contributes to improved report quality, especially in terms of temporal coherence and contextualization. The identified error modes highlight areas for future improvement, particularly in enhancing the model's ability to discern subtle visual cues and provide more precise quantitative descriptions, suggesting further integration with fine-grained visual reasoning and measurement extraction modules."

5. Conclusion

In this paper, we introduced **TempSem-GraphNet**, a novel framework designed to generate coherent and accurate radiology reports from multi-temporal Chest X-ray images, addressing limitations of existing methods that often lack historical context. Our core innovation is the *Multi-modal Temporal-Semantic Graph* (TempSem-Graph), which unifies visual lesion findings and semantic concepts from current and historical CXRs and reports. A **Hierarchical Temporal-Semantic Graph Attention Network (HTGAT)** processes this rich graph, integrating features into a fine-tuned LLaMA2-7B Large Language Model (LLM) for report generation. Experiments on MIMIC-CXR-JPG demonstrated TempSem-GraphNet's superior performance over strong baselines across all standard Natural Language Generation (NLG) and Clinical Entity (CE) metrics (e.g., BLEU-4 of 0.2441, CE F1-score of 0.288). A comprehensive ablation study confirmed the critical contribution of each proposed component, while human evaluations by board-certified radiologists rated TempSem-GraphNet's reports significantly higher in clinical accuracy, temporal coherence, and overall utility. In conclusion, TempSem-GraphNet represents a significant advancement in automated medical report generation, providing a powerful tool to enhance radiological practice and patient care through more comprehensive, context-rich information. Future work will address subtle finding omissions and integrate additional patient data modalities.

References

[1] Jiancheng Huang, Donghao Zhou, Jianzhuang Liu, Linxiao Shi, and Shifeng Chen. Ifast: Weakly supervised interpretable face anti-spoofing from single-shot binocular nir images. *IEEE Transactions on Information Forensics and Security*, 2024.

[2] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304. Association for Computational Linguistics, 2021.

[3] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical 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 (Volume 1: Long Papers)*, pages 3001–3012. Association for Computational Linguistics, 2021.

[4] Zixuan Zhou, Maycon Leone de Melo, and Tatiane Araújo Rios. Toward multimodal agent intelligence: Perception, reasoning, generation and interaction. 2025.

[5] Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10833–10842, 2024.

[6] Donghao Zhou, Jiancheng Huang, Jinbin Bai, Jiaze Wang, Hao Chen, Guangyong Chen, Xiaowei Hu, and Pheng-Ann Heng. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370*, 2024.

[7] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986v3*, 2021.

[8] Kaiyue Liu, Qi Sun, Daming Sun, Mengduo Yang, and Nizhuan Wang. Underwater target detection based on improved yolov7. *arXiv preprint arXiv:2302.06939v1*, 2023.

[9] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeffrey Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Veselin Stoyanov. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732. Association for Computational Linguistics, 2022.

[10] Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*, 2023.

[12] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. *arXiv preprint arXiv:2411.14961v3*, 2024.

[13] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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 (Volume 1: Long Papers)*, pages 5904–5914. Association for Computational Linguistics, 2021.

[14] Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P. Lungren, and Oge Marques. Structured dataset documentation: a datasheet for chexpert. *arXiv preprint arXiv:2105.03020v1*, 2021.

[15] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239. Association for Computational Linguistics, 2021.

[16] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513. Association for Computational Linguistics, 2021.

[17] Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924. Association for Computational Linguistics, 2022.

[18] Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123. Association for Computational Linguistics, 2021.

[19] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771. Association for Computational Linguistics, 2021.

[20] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007. Association for Computational Linguistics, 2021.

[21] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.

[22] Yucheng Zhou, Jihai Zhang, Guanjie Chen, Jianbing Shen, and Yu Cheng. Less is more: Vision representation compression for efficient video generation with large language models, 2024.

[23] Fangzhou Lin, Haotian Liu, Haoying Zhou, Songlin Hou, Kazunori D Yamada, Gregory S Fischer, Yanhua Li, Haichong K Zhang, and Ziming Zhang. Loss distillation via gradient matching for point cloud completion with weighted chamfer distance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 511–518. IEEE, 2024.

[24] Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956, 2024.

[25] Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, 2025.

[26] Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9262–9281, 2025.

[27] Yucheng Zhou, Lingran Song, and Jianbing Shen. MAM: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25319–25333, Vienna, Austria, July 2025. Association for Computational Linguistics.

[28] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911. Association for Computational Linguistics, 2023.

[29] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023. Association for Computational Linguistics, 2022.

[30] Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 27–38. Association for Computational Linguistics, 2021.

[31] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953. Association for Computational Linguistics, 2021.

[32] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405. Association for Computational Linguistics, 2021.

[33] Yonghao Liu, Renchu Guan, Fausto Giunchiglia, Yanchun Liang, and Xiaoyue Feng. Deep attention diffusion graph neural networks for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8142–8152. Association for Computational Linguistics, 2021.

[34] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics, 2024.

[35] Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958. Association for Computational Linguistics, 2022.

[36] Paul Röttger and Janet Pierrehumbert. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412. Association for Computational Linguistics, 2021.

[37] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021. Association for Computational Linguistics, 2021.