Spatio-Temporal and Retrieval-Augmented Modeling for Chest X-Ray Report Generation

Yan Yang, *Member, IEEE*, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, *Fellow, IEEE*, Weidong Han, and Jun Yu, *Senior Member, IEEE*

***Abstract*— Chest X-ray report generation has attracted increasing research attention. However, most existing methods neglect the temporal information and typically generate reports conditioned on a fixed number of images. In this paper, we propose STREAM: Spatio-Temporal and REtrieval-Augmented Modelling for automatic chest X-ray report generation. It mimics clinical diagnosis by integrating current and historical studies to interpret the present condition (temporal), with each study containing images from multi-views (spatial). Concretely, our STREAM is built upon an encoder-decoder architecture, utilizing a large language model (LLM) as the decoder. Overall, spatio-temporal visual dynamics are packed as visual prompts and regional semantic entities are retrieved as textual prompts. First, a token packer is proposed to capture condensed spatio-temporal visual dynamics, enabling the flexible fusion of images from current and historical studies. Second, to augment the generation with existing knowledge and regional details, a progressive semantic retriever is proposed to retrieve semantic entities from a preconstructed knowledge bank as heuristic text prompts. The knowledge bank is constructed to encapsulate anatomical chest X-ray knowledge into structured entities, each linked to a specific chest region. Extensive experiments on public datasets have shown the state-of-the-art performance of our method. Related codes and the knowledge bank are available at https://github.com/yangyan22/STREAM.**

Received 27 January 2025; revised 6 March 2025; accepted 21 March 2025. Date of publication 25 March 2025; date of current version 2 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62406093, Grant 62125201, Grant U24B20174, Grant 62422204, and Grant 62286082; in part by Zhejiang Provincial Natural Science Foundation of China under Grant LQ24F020032 and Grant LDT23F02025F02; and in part by the Central Government Guiding Fund for Local Science and Technology Development under Project 2024Y01018. *(Corresponding authors: Weidong Han; Jun Yu.)*

Yan Yang, Xiaoxing You, Ke Zhang, Jiajun Ding, Jiamei Sun, and Zhou Yu are with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yangyan@hdu.edu.cn; youxiaoxing@hdu.edu.cn; ke.zhang@hdu.edu.cn; djj@hdu.edu.cn; sunjiamei@hdu.edu.cn; yuz@hdu.edu.cn).

Zhenqi Fu is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: fuzhenqi@mail.tsinghua.edu.cn).

Xianyun Wang and Jun Yu are with the School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: wangxianyun@gmail.com; yujun@hit.edu.cn).

Qingming Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qmhuang@ucas.ac.cn).

Weidong Han is with the Department of Colorectal Medical Oncology, Zhejiang Cancer Hospital, Hangzhou 310022, China (e-mail: hanwd@zju.edu.cn).

Digital Object Identifier 10.1109/TMI.2025.3554498

***Index Terms*— Medical report generation, vision and language, spatio-temporal modeling, retrieval-augmented generation, large language models.**

I. Introduction

CHEST X-ray (CXR) plays a crucial role in clinical practice for diagnosing and monitoring the conditions of chest organs. Recently, automated CXR report generation has attracted significant research attention and made substantial progresses, aiming to assist radiologists by providing reference reports. However, current methodologies predominantly generate CXR reports based on a fixed number of images (e.g., a single image), neglecting temporal dynamics and multiview information inherent in clinical practice.

In real-world scenarios, radiologists combine multiview spatial visual information and inspect anatomical chest regions individually to diagnose a given study (referred to as "spatial analysis"). Typically, a study encompasses CXR images captured from diverse perspectives, including posteroanterior (PA), lateral (L) or anteroposterior (AP) views. Additionally, radiologists routinely compare the current study with historical studies to acquire the progression of a medical condition (referred to as "temporal analysis"). As illustrated in Fig. 1, the clinical diagnosis integrates both spatial and temporal information, ensuring a more accurate and comprehensive diagnosis. This allows for the diagnosis of disease progressions such as "larger" and "stable", providing detailed diagnosis information about the changes over time.

Existing methods, such as [1], [2], [3], [4], and [5], generate the report based on a single image on the MIMIC-CXR dataset [6]. These methods display inaccuracies in depicting disease progression, and generate progressions by the model's data bias and hallucinations. Inspired by the diagnostic workflows of radiologists, several prior works have explored incorporating either temporal or multiview spatial information for CXR report generation. For example, previous works [7], [8] propose integrating images from the current study and one historical study, along with the historical report, to generate the current report. However, these methods overlook the importance of multiview information within each study. In contrast, another study [9] focuses on multiview image integration for report generation but does not account for the temporal dynamics.

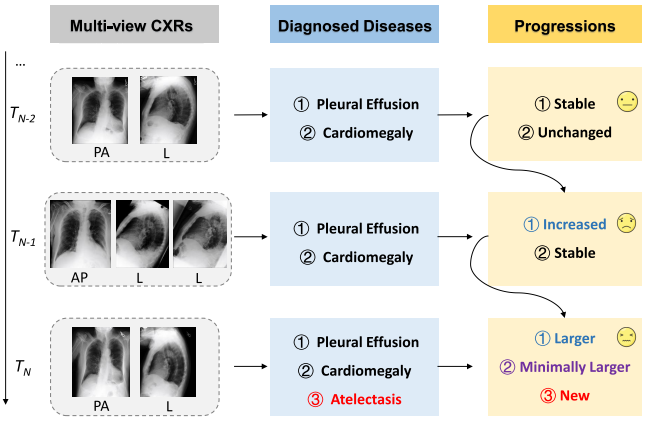

Fig. 1. An example of spatio-temporal analysis in clinical diagnosis. The progression of diseases are acquired by comparing the current study with historical studies. Additionally, multiple views, such as posteroanterior (PA) and lateral (L) images in a single study, are integrated to aid in diagnosis. T_i denotes the i_{th} study of the patient.

In this paper, we depart from prior approaches by integrating multiview spatio-temporal dynamic information and performing region-level semantic retrieval to enhance automated CXR report generation, closely mimicking the clinical diagnostic process. Specifically, we introduce STREAM, a model that combines spatio-temporal and retrieval-augmented techniques to achieve accurate diagnoses. It utilizes a large language model (LLM) as the decoder, incorporating spatio-temporal visual dynamics as visual prompts and predicting regional semantic entities as textual prompts. First, our proposed token packer effectively captures spatio-temporal visual dynamics from a flexible array of images, including multiview images from both current and historical studies. Second, to emulate the clinical diagnostic process of examining regions in sequence, we introduce a progressive retriever that extracts region-level semantic entities from our pre-constructed cross-modal knowledge bank. These retrieved entities act as heuristic textual prompts, providing vital anatomical knowledge of chest X-rays that significantly augments the generation process. In summary, our key contributions are listed as follows:

- We propose STREAM that performs spatio-temporal and retrieve-augmented modelling for CXR report generation. An LLM is implemented as the report generator, taking spatio-temporal visual dynamics as visual prompts and regional semantic entities as textual prompts.
- We propose a token packer to capture condensed visual dynamics from a flexible number of images, and a progressive semantic retriever to derive regional semantic entities from our preconstructed knowledge bank to augment the generation. The knowledge bank encapsulates CXR anatomical knowledge into structured entities.
- Extensive experiments on public datasets, evaluated by various metrics, have demonstrated the state-of-the-art performance of our method. To the best of our knowledge, our STREAM is the first work trying to capture multiview spatio-temporal dynamics and conduct region-level semantic retrieval for CXR report generation.

II. Related Works

In this section, we review related works on chest X-ray report generation, spatio-temporal modelling, and retrieval-augmented generation.

A. Chest X-Ray Report Generation

Chest X-ray (CXR) report generation has received increasing research attention. RGRG [1] proposes to detect anatomical regions and then describe all regions to form the complete report. FMVP [9] proposes a flexible multiview paradigm for generation using an observation-to-concept workflow. R2Gen [5], CMN [10], CMM [11] and MA [12] propose to integrate memory networks and learn the vision-language alignment for CXR report generation. Clinical-BERT [2] learns image-text-MeSH alignment via the BERT-like pre-training. RAMT [13] proposes a relation-aware mean teacher method for semi-supervised generation. ORGAN [14] proposes observation-guided radiology report generation via tree reasoning. Token-Mixer [15] proposes learning cross-modal alignment by a token-mixing strategy and the alternative training. DCL [4] proposes a dynamic graph to enhance medical contrastive learning [16] for report generation. PromptMRG [17] proposes to guide the generation with diagnosis-aware prompts using an extra disease classification branch. It also retrieves similar reports from the database to assist the diagnosis of a query image. METransformer [3] proposes learnable expert tokens and simulates the multi-expert joint diagnosis to determine the final report. GSK [18] proposes employing the general and specific knowledge to enhance the CXR report generation. AP-ISG [19] proposes a region-based attribute prototype-guided iterative scene graph for explainable report generation. Xu et al. [20] constructed a RadioGraphy Captions dataset and pre-trained a vision-language Transformer for report generation. Recently, there has been a trend towards integrating LLMs [21] for CXR report generation. CheX-agent [22] proposes an instruction-tuned foundation model for CXR interpretation. R2GenGPT [23] proposes efficient visual alignment, aligning the vision and language space. MAIRA-1 [24] leverages a CXR-specific image encoder in combination with a fine-tuned LLM and text-based data augmentation, to generate reports. LLM-CXR [25] proposes an instruction-finetuning for LLMs, enabling the alignment between images and texts for multimodal CXR tasks.

B. Spatio-Temporal Modelling

Spatio-temporal modelling is crucial for the comprehensive analysis of temporal visual data, including video analysis and dynamic medical image analysis. Video-LLaMA [26] introduces a multimodal framework that equips LLMs with the ability to understand video content. The framework utilizes a frozen pre-trained image encoder to extract features from video frames, incorporates a position embedding layer to embed temporal information into the frames, and employs a video Q-Former [27] to aggregate frame-level representations. In the medical domain, various spatio-temporal models have been developed, facilitating the monitoring of disease progression, assessment of treatment responses, and prediction

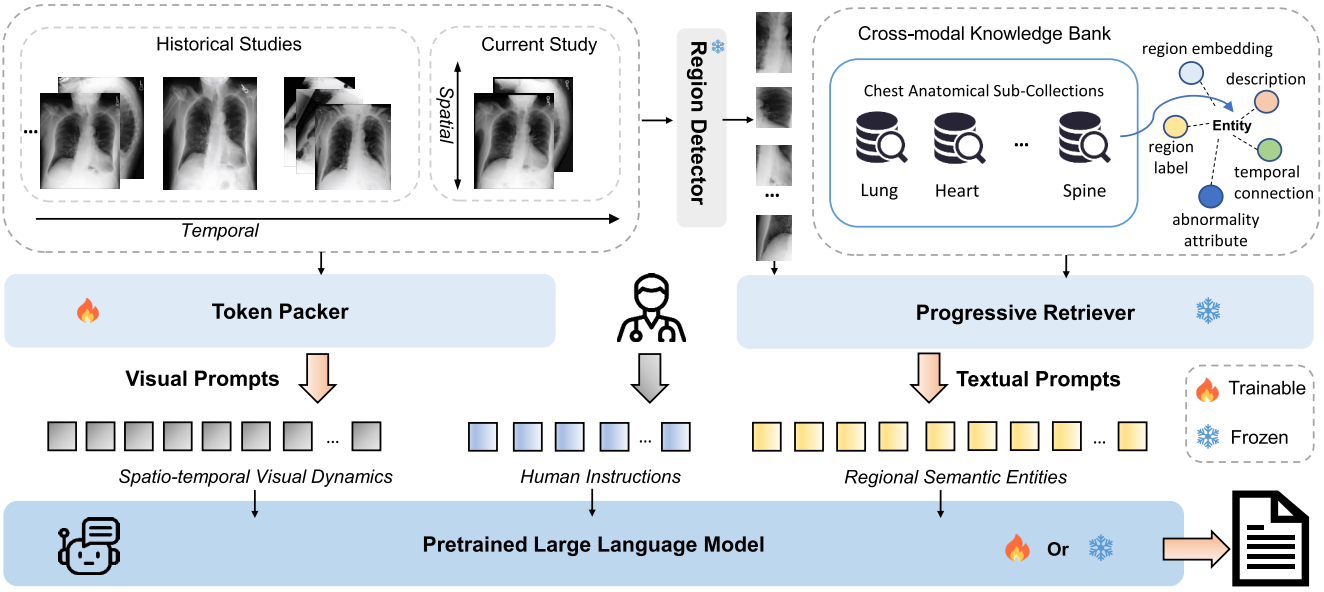

Fig. 2. Overall framework of our STREAM. A large language model is utilized as the decoder, taking spatio-temporal visual dynamics (captured by a token packer), fine-grained regional semantic entities (predicted from a preconstructed cross-modal knowledge bank) and human instructions (manually defined to prompt the large language model) as inputs for CXR report generation. Entities in the knowledge bank are represented as quintuples, each comprising a description, an embedding, a label, an abnormality attribute, and the temporal connection of a specific region.

of patient outcomes. Zhu et al. [28] propose a spatio-temporal graph hubness propagation model for dynamic brain network classification. Ahmadi et al. [29] propose Transformer-based spatio-temporal analysis for classifying aortic stenosis severity from echocardiography cine series. TransVFS [30] proposes a spatio-temporal local–global Transformer to extract local image details and the global dependency to learn prostate deformations for force estimation. BioViL-T [31] proposes a vision-language pretraining method, effectively learning the alignment between the report and temporal images. RECAP [8] proposes to first predict observations and progressions for two consecutive images, and then integrate the historical report, spatio-temporal information, and the current image for the current report generation. HERGen [32] proposes to employ a group causal Transformer to integrate longitudinal data across patient visits for report generation. Mei et al. [33] proposes to integrate historical and current patient data (including both prior images and reports) for medical report generation, which tackles textual diversity in cross-modal tasks through style-agnostic representations and advanced token prediction. In this paper, we propose STREAM, the first spatio-temporal model that integrates dynamic temporal and multiview spatial information for automated CXR report generation.

C. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) [34] has emerged as a powerful technique to promote the capabilities of LLMs by integrating external knowledge. RAG helps maintain up-to-date knowledge, incorporate long-tail knowledge [35], and protect private training data [36]. A typical RAG process involves the following steps. Given an input query, the retriever locates and looks up relevant data sources, then the retrieved results will interact with the generator to guide the generation process. EchoSight [37] proposes a multimodal RAG framework that enables LLMs to answer visual questions which require fine-grained encyclopedic knowledge. Sarto et al. [38] propose the retrieval-augmented Transformer for image captioning that integrates KNN-augmented attention layers to generate words based on textual sentences retrieved from an external memory. SmallCap [39] proposes to generate a caption conditioned on an input image and related captions retrieved from a database. TranSQ [40] retrieves sentences for the query CXR via a sentence retriever and a sentence selection module. In this paper, we propose retrieving regional semantic entities from our pre-constructed knowledge bank for detected CXR regions as textual prompts to augment report generation.

III. Method

The proposed STREAM generates the current diagnostic report conditioned on image sequences from both the current and historical studies. As shown in Fig. 2, STREAM incorporates a pretrained LLM as the decoder, leveraging spatio-temporal visual dynamics as visual prompts and regional semantic entities as textual prompts. Initially, a token packer captures spatio-temporal visual dynamics by flexibly packing a variable number of images (from current and historical studies) into condensed visual tokens. Next, a cross-modal knowledge bank with multiple anatomical sub-collections is established, each sub-collection contains structured semantic entities of a specific chest region. Following this, regions of the target CXR image are detected by a pre-trained detector. These detected regions are then attached with regional semantic entities by a progressive semantic retriever to augment the generation.

A. Spatio-Temporal Visual Dynamic Extraction

Overall, spatio-temporal visual dynamics are captured by a token packer. With the advent of large pre-trained models, such

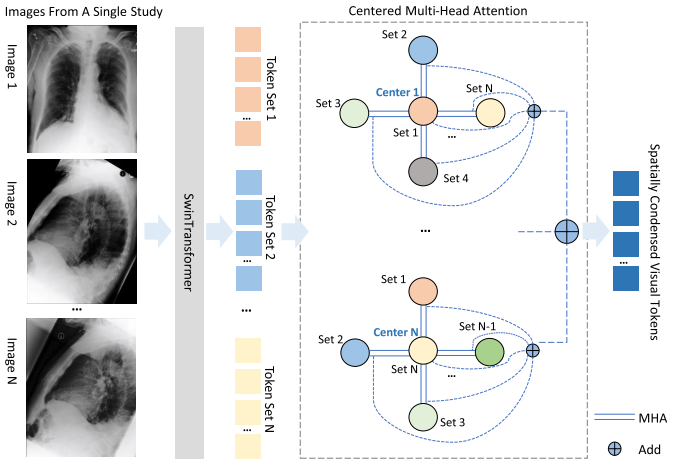

Fig. 3. The acquisition of spatially condensed visual tokens. Image token sets from a single study are initially encoded by SwinTransformer. Then, Centered multi-head attention (Centered MHA) is applied across all image token sets, enabling integration among tokens from different views. In this process, image token sets from different views are alternatively selected as the center (i.e., query in the MHA) and the rest token sets are defined as the keys and values.

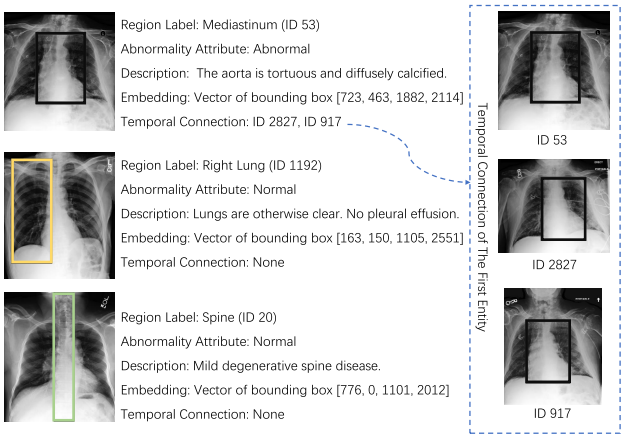

Fig. 4. Examples of entities within the cross-modal knowledge bank. Each entity comprises five key elements: a region label, an abnormality attribute, a description, an embedding, and a temporal connection. As depicted in the first entity, the temporal connection indicates the link between the current entity and its historical counterparts.

as LLMs, capable of processing various tasks, effective token compression has become increasingly critical. In this paper, we need to handle a large and variable number of multi-view images from both current and historical studies. To this end, we introduce a token packer designed to efficiently fuse image tokens from various views and studies. Below, we explain the process of how information from different views and studies is handled.

Initially, images from both current and historical studies are divided into patches and processed by a SwinTransformer [41] to generate image tokens, which can be formulated as:

$$E = \{e_1, e_2, \ldots, e_{N_P}\} = \text{SwinTransformer}\,(X)\,, \qquad (1)$$

where E represents the set of image tokens from the input image X, and e_i is the i_{th} patch embedding of the input image. Here, N_P denotes the number of patch embeddings. Rather than directly concatenating all image tokens as visual prompts, we fuse image tokens from multiple views and studies into compact tokens. This effectively reduces token length, mitigating the redundancy and inefficiency.

As illustrated in Fig. 3, we first integrate images from different views within a single study using the proposed Centered Multi-Head Attention (Centered MHA) to generate spatially condensed visual tokens. The process is formulated as follows:

$$S = \sum_{i=1}^{N}\Big(\sum_{j=1,\ j\neq i}^{N} \text{MHA}(E^i, E^j, E^j) + E^i\Big), \qquad (2)$$

where E^i represents the set of image tokens for the i_{th} image X^i, serving as the query (i.e., center) in our Centered MHA [42]. E^j is the set of image tokens for the j_{th} image X^j, serving as the key and value of the MHA. S denotes the spatially condensed visual tokens, N is the number of images in a study. As shown in Fig. 3, image token sets from different views are alternatively selected as the center (i.e., query in the MHA) and the rest token sets are defined as the keys and values. Then, the outputs of the MHA from each pair of token sets surrounding the center are aggregated through summation. Following this, in our Centered MHA, we implement a residual connection (which turns out to be important), where the center token set are added to the aggregated output. Subsequently, the aggregation results from all central sets are summed to produce spatially condensed visual tokens.

Finally, spatially condensed visual tokens from all studies will be integrated to generate the final spatio-temporal visual dynamics. This process is formulated as:

$$\text{Prompt}_{\text{vision}} = \left[S^c; S^{h_1} + S^{h_2} \ldots + S^{h_H}\right], \qquad (3)$$

where S^c denotes the spatially condensed visual tokens of the current study. S^{h_i} denotes the spatially condensed visual tokens of the i_{th} historical study, H denotes the number of historical studies. [;] denotes the concatenation operation. $\text{Prompt}_{\text{vision}}$ refers to the spatio-temporal visual dynamics which will be fed to the LLM as visual prompts.

B. Cross-Modal Knowledge Bank Construction

This section details our approach for building a cross-modal knowledge bank that encapsulates chest anatomical knowledge into semantic entities, each linked to a specific CXR region. The entity examples are illustrated in Fig. 4. Specifically, each entity contains a region label, an abnormality attribute (indicating whether the region is normal or abnormal), a description of the region, an embedding of the region given the bounding box, and a temporal connection (linking identical regions across different studies of the same patient).

The construction of the knowledge bank involves the following key steps: 1) Meta-Data Extraction: The region label, abnormality attribute, description and bounding box of each CXR region are derived from the Chest ImaGenome dataset [43] using the related codes.[1] Chest ImaGenome

[1]https://github.com/ttanida/rgrg

decomposes CXR images from the MIMIC-CXR dataset [6] into 29 anatomical regions, automatically annotating each region with related descriptions derived from the paired radiology reports. 2) Anatomical Region Selection: We select anatomical regions by eliminating overlapping regions and prioritizing the most clinically significant ones. For example, we exclude the "upper mediastinum" and select the broader "mediastinum" region instead. 3) Embedding and Temporal Connection Generation: The region within bounding boxes are encoded using the image encoder of MedCLIP [44] to generate region embeddings. Then, each entity is linked to its temporal counterparts based on the metadata of MIMIC-CXR. 4) Entity Selection: We select entities based on region descriptions to ensure that most diseases are included in the knowledge bank. Besides, since the region description is automatically annotated, we further filter out entities with inaccurate or noisy descriptions to maintain the quality and accuracy. 5) Bounding Box Validation: Entities with incorrectly labelled bounding boxes are manually identified and deleted via visual inspection of the images. It is worth noting that all entities are sourced from the training datasets of both MIMIC-CXR and Chest ImaGenome to prevent information leakage during testing. Finally, each structured entity in the knowledge bank is represented as:

$$entity_i = \{v_i, l_i, s_i, a_i, c_i\}, \tag{4}$$

where the entity contains five elements: a region embedding v_i captured by MedCLIP's visual encoder, a region label l_i identifying the anatomical organ (e.g., "spine"), a description s_i detailing the condition of the region, an abnormality attribute a_i distinguishing the normal and abnormal states, and a temporal connection c_i indicating the link between the current entity and its historical counterparts. Importantly, we establish temporal connections among entities with the same region label across different studies of a single patient, enhancing the understanding of region-level temporal dynamics. The structure enables the knowledge bank to function as a comprehensive repository including anatomical cross-modal knowledge, enriched with temporal dynamics. To facilitate efficient retrieval, the knowledge bank is divided into sub-collections, grouping entities with region labels. In summary, the knowledge bank is a repository containing sub-collections that integrate region embeddings, labels, textual descriptions, abnormality attributes, and temporal connections. This structure offers a comprehensive view of anatomical regions and their temporal conditions.

Regarding the maintenance of the knowledge bank, we focus on two key aspects: 1) Regular Updates: We plan to regularly update the knowledge bank based on the latest medical literature, clinical cases, and research findings, ensuring that it incorporates typical and the most updated CXR anatomical entities. 2) User Feedback Mechanism: We plan to collect feedback from users to identify and address any potential issues, ensuring that the knowledge bank aligns with researchers' needs, such as correcting descriptions of entities and expanding the coverage of anatomical regions.

C. Progressive Semantic Retrieval

To enrich the generation process with detailed and region-specific information, we propose retrieving region-level semantic entities for CXR images from the preconstructed knowledge bank. For a given CXR image, we first detect regions with a pre-trained Faster R-CNN [45]. Next, we employ MedCLIP to extract visual embeddings for these detected regions. Each region is then attached with a corresponding semantic entity by our progressive retrieval. Specifically, for a detected query region, we compute the cosine similarity between the query region's embedding and the embeddings of regions from the knowledge bank's sub-collection. The similarity score is calculated as:

$$similarity\left(v_q, v_j\right) = < \frac{v_q \cdot v_j}{\|v_q\| \cdot \|v_j\|} >, j \in 1, 2, \ldots, N_S, \tag{5}$$

where v_q is the visual embedding of the query region and v_j is the visual embedding of the j_{th} entity in the sub-collection. N_S denotes the number of entities in the sub-collection. We employ the FAISS library [46] for efficient vector search.

The progressive retrieval involves selecting candidate entities (as depicted in Algorithm 1) and then using a voting mechanism to determine the best match (denoted as the retrieved entity) for each query region. Initially, given a query region, candidate entities with the top K similarity scores are chosen from the sub-collection. Then, the progressive retrieval iterates for several time steps. At each time step, if historical regions for the query and candidate entities are available, we calculate historical similarity scores between the query's historical region and those of the candidate entities. Candidate entities are then updated by retaining those with similarity scores above a predefined threshold. If no historical regions are available or if similarity scores fall below the threshold, the candidate entities from the previous time step are retained.

Algorithm 1 Selection of Candidate Entities

Input: Embeddings of the query region v_q and its associated historical regions $\{v_q^1, v_q^2, \ldots, v_q^T\}$ over time steps T.
Output: Candidate entities.

Initialize candidate entities with region embeddings $\{v_1, v_2, \ldots, v_K\}$ that exhibit the top K similarity scores.
Set $t = 0$.
while $t < T$ **do**
  Increment t by 1.
  if historical regions of candidate entities exist at time t **then**
    Compute similarity scores between historical region embeddings v_q^t and $\{v_1^t, v_2^t, \ldots, v_K^t\}$.
    Update candidate entities with those having historical similarity scores above a predefined threshold.
  else
    Retain candidate entities from the time step $t - 1$.
  end if
end while
return Candidate Entities.

Subsequently, a voting process is conducted to determine abnormality of the query region by aggregating abnormality attributes of candidate entities. If the count of abnormal candidate entities exceeds that of normal entities, the query region is classified as abnormal. In this case, the abnormal candidate entity with the highest similarity score is selected as the retrieved entity. Conversely, if the majority of candidate entities are normal, the query region is deemed normal, and the entity with the highest similarity score among normal candidate entities is selected as the retrieved entity. Finally, for the target CXR image, each detected region is mapped to its retrieved entity. The descriptions of these entities are then concatenated and provided to the LLM as textual prompts:

$$\text{Prompt}_{\text{text}} = [\{l_1 : s_1\}; \{l_2 : s_2\}; \ldots; \{l_M : s_M\}], \quad (6)$$

where $\text{Prompt}_{\text{text}}$ is the textual prompt comprising region labels and their associated descriptions. [;] indicates the concatenation operation. $\{l_i : s_i\}$ denotes the label l_i and the description s_i of the i_{th} entity. The variable M denotes the number of entities included in the textual prompt.

D. Parameter Training

In summary, our STREAM incorporates the visual prompts (i.e., multiview spatio-temporal visual dynamics), the textual prompts (i.e., regional semantic entities) and the human instructions, as inputs to an LLM for report generation. The generation process can be formulated as the following auto-regressive equation:

$$p(R) \sim \prod_{t=1}^{L} p_\theta\left(r_t \mid \text{Prompt}_{\text{vision}}, \text{Prompt}_{\text{text}}, I, R_{0:t-1}\right), \quad (7)$$

where R represents the complete report sequence with a length of L. p denotes the probability distribution. p_θ denotes the probability distribution parameterized by θ. r_t is the t_{th} word in the report, while $R_{0:t-1}$ denotes the previous report sequence. $\text{Prompt}_{\text{vision}}$ denotes visual prompts. $\text{Prompt}_{\text{text}}$ stands for the textual prompts. I refers to the human instructions, which serves to establish a clear goal for the LLM, enabling it to grasp the task and the anticipated outputs (i.e., detailed chest X-ray report including descriptions of any abnormalities or significant findings, along with the view and temporal progression details). Our STREAM is trained by a cross-entropy loss function, formulated as:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{L} \log\left(p_\theta\left(r_t \mid \left(X^c, X^h, I, R_{0:t-1}\right)\right)\right), \quad (8)$$

where $\mathcal{L}_{\text{LM}}$ is the loss function of our STREAM. p_θ is the probability of generating word r_t at time step t conditioned on the previous report sequence $R_{0:t-1}$, the current image sequence X^c and historical image sequences $X^h = \{X^{h_1}, \ldots, X^{h_H}\}$, where H denotes the number of historical studies. θ is the network parameters. In our method, the historical image sequences can be empty $X^p = \emptyset$ if no historical study exists.

IV. Experiments

To evaluate the performance, we carry out extensive experiments on public datasets, including performance comparisons, ablation studies, case studies, human evaluations, etc.

A. Datasets and Evaluation Metrics

We evaluate the performance of our STREAM on the publicly available datasets: (1) MIMIC-CXR [6]: This dataset contains 377,110 chest X-ray images and 227,835 reports. Following previous studies [5], we use the official data split and remove samples without "findings". (2) IU X-Ray [47]: This dataset includes 7,470 images and 3,955 reports, with each report linked to either a single image or multiple images. Following previous studies [5], we exclude samples without "findings" and use a data split of 70% for training, 10% for validation, and 20% for testing.

To assess the performance, we utilize a set of natural language generation (NLG) metrics, including Bilingual Evaluation Understudy (BLEU) [48], Consensus-based Image Description Evaluation (CIDEr) [49], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [50], and METEOR [51]. We also use clinical efficacy (CE) metrics for evaluation, including the F1 score, Recall and Precision. Additionally, the RaTEScore [52], CheXbert vector similarity [53], and RadGraph [54] F1 are used to comprehensively evaluate the performance [55]. CheXbert vector similarity calculates the semantic embedding similarity between reports, while Radgraph F1 analyzes the consistency of key entities (e.g., anatomical structures and lesion types) and their relationships in the generated reports. CheXbert vector similarity and Radgraph F1 are tested using codes.[2] RaTEScore, focusing on crucial medical entities such as diagnostic outcomes and anatomical details, utilizes entity embeddings to compare the similarity and clinical relevance of extracted entities. It is robust against medical synonyms and sensitive to negation expressions. The evaluation codes are publicly available.[3] Besides, we conduct case studies and human evaluations on generated reports to further evaluate the performance.

B. Implementation Details

Our STREAM is trained in an end-to-end manner using PyTorch on four NVIDIA 3090 GPUs. Image tokens are extracted via a SwinTransformer (Swin-Base) [41], and region embeddings are derived from a pretrained ViT model in MedCLIP [44]. The large language model used in our method is the TinyLlama-1.1B [56]. For chest X-ray region detection, we utilize a Faster R-CNN model, following a training pipeline similar to RGRG [1]. The multi-head attention mechanism is with 8 heads, and the batch size is set to 8. We employ the AdamW optimizer [57] with a cosine learning rate schedule, starting with an initial learning rate of 1e-4. The training is limited to a maximum of 6 epochs. In the proposed progressive retriever, the hyper-parameter K is set to 50. For each study, we limit the maximum number of multiview images to 3. For

[2]https://github.com/rajpurkarlab/CXR-Report-Metric
[3]https://github.com/MAGIC-AI4Med/RaTEScore

TABLE I

PERFORMANCE COMPARISONS ON MIMIC-CXR CONCERNING NLG METRICS, RATESCORE, CHEXBERT VECTOR SIMILARITY (DENOTED AS CHEXBERT), AND RADGRAPH F1 (DENOTED AS RADGRAPH). † INDICATES THE RESULT WAS REPRODUCED. ‡ DENOTES OUR METHOD INTEGRATING A SINGLE VIEW IN EACH STUDY. * INDICATES THE METHODS TAKE HISTORICAL REPORTS AS INPUT FOR CURRENT REPORT GENERATION. THE BEST RESULTS ARE MARKED IN BOLD

Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	RaTEScore	CheXbert	RadGraph
R2Gen [5]	2020	0.353	0.218	0.145	0.103	0.142	0.277	0.478	0.331	0.176
CMN [10]	2021	0.353	0.218	0.148	0.106	0.142	0.278	0.488	0.342	0.174
CMM [11]	2022	0.381	0.232	0.155	0.109	0.151	0.287	-	-	-
GSK [18]	2022	0.363	0.228	0.156	0.115	-	0.284	-	-	-
Clinical-BERT [2]	2022	0.383	0.230	0.151	0.106	0.144	0.275	-	-	-
DCL [4]	2023	-	-	-	0.109	0.150	0.284	-	-	-
METransformer [3]	2023	0.386	0.250	0.169	0.124	0.152	0.291	-	-	-
R2GenGPT [23]†	2023	0.409	0.261	0.179	0.128	0.160	0.287	0.506	0.381	0.212
RECAP [8]*	2023	0.429	0.267	0.177	0.125	0.168	0.288	-	-	-
HERGen [32]	2024	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-
Mei *et al.* [33]*	2024	0.436	0.275	0.184	0.129	**0.177**	**0.305**	-	-	-
RAMT [13]	2024	0.362	0.229	0.157	0.113	0.153	0.284	-	-	-
FMVP [9]	2024	0.389	0.236	0.156	0.108	0.150	0.284	-	-	-
PromptMRG [17]	2024	0.398	-	-	0.112	0.157	0.268	0.479	**0.439**	0.177
MA [12]	2024	0.396	0.244	0.162	0.115	0.151	0.274	-	-	-
Our STREAM	2024	0.422	0.271	0.188	0.137	0.165	0.290	0.516	0.397	0.219
Our STREAM‡	2024	0.420	0.267	0.184	0.133	0.164	0.291	0.512	0.392	0.218
Our STREAM*	2024	**0.437**	**0.278**	**0.192**	**0.139**	0.172	0.297	**0.520**	0.405	**0.223**

the diagnosis of the current study, we constrain the maximum number of temporal studies to 3 (i.e., including a current study and two historical studies). In our implementation, the token packer is trainable, and the region detector is pretrained and frozen. The TinyLlama-1.1B can either be frozen or fine-tuned using LoRA [58]. Our cross-modal knowledge bank comprises 11 sub-collections, each containing over 3,000 representative structured entities. The anatomical regions in the sub-collections are "right lung", "right hilar structures", "right apical zone", "left lung", "left hilar structures", "left apical zone", "trachea", "spine", "mediastinum", "cardiac silhouette", and "abdomen". We set the human instruction to "Please generate a detailed chest X-ray report, including descriptions of any abnormalities or significant findings, along with the view and temporal progression details" to prompt the LLM.

C. Performance Comparisons

We compare our STREAM against 15 state-of-the-art methods from recent literatures, including R2Gen [5], CMN [10], CMM [11], GSK [18], Clinical-BERT [2], DCL [4], METransformer [3], RAMT [13], FMVP [9], MA [12], R2GenGPT [23], RECAP [8], HERGen [32], Mei et al. [33], and PromptMRG [17]. All of these methods follow an encoder-decoder pipeline. For the image encoder, R2Gen, CMN, CMM, GSK and PromptMRG use ResNet-101 [59], while Clinical-BERT, RAMT, FMVP, Mei et al., and MA employ DenseNet-121 [60]. In contrast, DCL, METransformer, R2GenGPT, RECAP, HERGen and our STREAM leverage Vision Transformer [61] or its variants, such as SwinTransformer [41], as the vision encoder. Regarding the decoder, most of the comparison methods use Transformer or its variants as the decoder, while R2GenGPT and our STREAM utilize LLMs for decoding. Specifically, R2GenGPT employs the Llama-2-7B as the decoder and our STREAM utilizes the TinyLlama-1.1B as the decoder. Among the methods, FMVP and Mei et al. take multiview CXR images as input. RECAP, HERGen and Mei et al. integrate the temporal study for generation. It is noted that we reproduced R2GenGPT for comparison, since it was originally trained and tested with the "impression" section. For the other methods, the results were directly taken from their respective papers or reproduced by testing the released models. Details of the methods are provided in the introduction section. Moreover, similar as Mei et al. and RECAP, we also test the performance when the historical prior report is integrated to our STREAM on MIMIC-CXR. Additionally, we assess the performance of our STREAM model when a single view per study is used for report generation.

Performance comparisons (concerning NLG metrics, RaTEScore, CheXbert vector similarity, and Radgraph F1) on MIMIC-CXR and IU X-Ray are presented in Table I and Table II respectively, with the best results highlighted in bold. As can be observed, our STREAM demonstrates remarkable performance across most of the metrics on both datasets. In particular, our STREAM achieves a significant improvement in BLEU-3 and BLEU-4 scores on MIMIC-CXR, outperforming the existing methods by a large margin. Additionally, our STREAM attains the highest BLEU-1, BLEU-2, BLEU-3, BLEU-4 and METEOR scores on IU X-Ray. Furthermore, our method achieves the competitive METEOR and ROUGE-L scores on both datasets.

We also assess the performance using RaTEScore, CheXbert vector similarity, RadGraph F1. For comparison, we test the released models for R2Gen, CMM, and PromptMRG, and reproduce R2GenGPT. As shown, our method achieves the highest RaTEScore and RadGraph F1 on both datasets. PromptMRG achieves the best CheXbert vector similarity, as it includes an additional classification branch tailored to enhance the CE accuracy. However, it is worth noting that PromptMRG uses a model trained on MIMIC-CXR to test on IU X-Ray, which leads to suboptimal performance on the latter dataset. Following the previous methods [5], we further assess our STREAM's performance using CE metrics on MIMIC-CXR over 14 diseases. We directly assign classification labels for the

TABLE II
PERFORMANCE COMPARISONS ON IU X-RAY CONCERNING NLG METRICS, RATESCORE, CHEXBERT VECTOR SIMILARITY (ABBREVIATED AS CHEXBERT), AND RADGRAPH F1 (ABBREVIATED AS RADGRAPH). † INDICATES THAT THE METHOD WAS REPRODUCED. ‡ DENOTES OUR METHOD INTEGRATING A SINGLE VIEW IN EACH STUDY. THE BEST RESULTS ARE MARKED IN BOLD

Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	RaTEScore	CheXbert	RadGraph
R2Gen [5]	2020	0.470	0.304	0.219	0.165	0.187	0.371	0.615	0.530	0.318
CMN [10]	2021	0.475	0.309	0.222	0.170	0.191	0.375	0.603	0.552	0.295
CMM [11]	2022	0.494	0.321	0.235	0.181	0.201	0.384	-	-	-
GSK [18]	2022	0.496	0.327	0.238	0.178	-	0.381	-	-	-
Clinical-BERT [2]	2022	0.495	0.330	0.231	0.170	0.209	0.376	-	-	-
DCL [4]	2023	-	-	-	0.163	0.193	0.383	-	-	-
METransformer [3]	2023	0.483	0.332	0.228	0.172	0.192	0.380	-	-	-
R2GenGPT [23]†	2023	0.491	0.323	0.234	0.180	0.211	0.376	0.641	0.601	0.321
RAMT [13]	2024	0.482	0.310	0.221	0.165	0.195	0.377	-	-	-
FMVP [9]	2024	0.485	0.315	0.225	0.169	0.201	**0.398**	-	-	-
PromptMRG [17]	2024	0.401	-	-	0.098	0.160	0.281	0.577	0.556	0.249
MA [12]	2024	0.501	0.328	0.230	0.170	0.213	0.386	-	-	-
Our STREAM	2024	**0.506**	**0.338**	**0.248**	**0.188**	**0.215**	0.387	**0.659**	**0.615**	**0.357**
Our STREAM‡	2024	0.499	0.333	0.238	0.178	0.213	0.377	0.652	0.608	0.330

TABLE III
PERFORMANCE COMPARISONS CONCERNING CE METRICS ON MIMIC-CXR. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. † INDICATES THE METHOD WAS REPRODUCED. * INDICATES METHODS INTEGRATING HISTORICAL REPORT FOR GENERATION. ‡ DENOTES OUR METHOD INTEGRATING ONLY A SINGLE VIEW IN EACH STUDY

Method	Year	Precision	Recall	F1
R2Gen [5]	2020	0.333	0.273	0.276
CMN [10]	2021	0.344	0.275	0.278
CMM [11]	2022	0.342	0.294	0.292
GSK [18]	2022	0.458	0.348	0.371
Clinical-BERT [2]	2022	0.397	0.435	0.415
DCL [4]	2023	0.471	0.352	0.373
METransformer [3]	2023	0.364	0.309	0.311
R2GenGPT [23]†	2023	0.485	0.383	0.428
RECAP [8]*	2023	0.389	0.443	0.393
HERGen [32]	2024	0.415	0.301	0.317
Mei *et al.* [33]*	2024	-	-	0.411
RAMT [13]	2024	0.380	0.342	0.335
FMVP [9]	2024	0.332	0.383	0.336
PromptMRG [17]	2024	0.501	**0.509**	0.476
MA [12]	2024	0.411	0.398	0.389
Our STREAM	2024	0.527	0.412	0.462
Our STREAM‡	2024	**0.531**	0.407	0.460
Our STREAM*	2024	0.515	0.447	**0.478**

reference and the generated reports to calculate clinical efficacy scores. This process leverages the CheXbert labeler [53], which assigns one of four tags to each disease: "−1: uncertain", "0: negative", "1: positive", and "None: not mentioned". We further set "None" to "0" and "−1" to "1". Then, we compare the labels assigned to the generated reports with those of the reference reports to determine the accuracy scores. The results are shown in Table III. As can be observed, our method attains the best precision, while maintaining a competitive performance in Recall scores, comparable to the current state-of-the-art methods. Notably, the method PromptMRG incorporates an extra classification branch tailored to enhance the clinical accuracy, and thus yields the highest recall among the comparison methods regarding the 14 predefined diseases. However, as can be observed from Table I and Table II, PromptMRG demonstrates average performance concerning the NLG metrics. Besides, as shown in Table I, and Table III, the integration of prior reports improves the performance of our STREAM model, as many disease descriptions remain unchanged over time. Furthermore, multi-view images lead to better performance compared with single-view integration. Overall, our STREAM method demonstrates excellent performance across NLG metrics, RaTEScore, and RadGraph, while maintaining competitive results in CE metrics and CheXbert vector similarity. To the best of our knowledge, our STREAM is the first work trying to capture multiview spatio-temporal dynamics and conduct region-level semantic retrieval for CXR report generation.

D. Ablation Studies

We also conduct ablation studies to assess individual contributions of the components within our STREAM. Ablation results on MIMIC-CXR and IU X-Ray are presented in Table IV and Table V, respectively. Concretely, we test the components "Multiview Integration", "Temporal Incorporation" and "Semantic Entity Retrieval" on MIMIC-CXR, and test the components "Multiview Integration" and "Semantic Entity Retrieval" on IU X-Ray. The baseline model employs a single image from the current study for generation. "Multiview" (i.e., Multiview Integration) integrates multiview images from the current study, "Temporal" (i.e., Temporal Incorporation) incorporates temporal data from historical studies, and "Entity" (i.e., Semantic Entity Retrieval) integrates retrieved semantic entities as textual prompts to enhance CXR report generation.

As illustrated in Table IV and Table V, both "w. multiview" and "w. entity" outperform "baseline", highlighting the efficacy of multiview integration and regional entity retrieval. In Table IV, "w. temporal" outperforms the baseline, demonstrating the efficacy of integrating the temporal information into generation. Besides, in Table IV, our STREAM with all components outperforms the "w. multiview and temporal". Similarly, "w. entity" outperforms the baseline model, and "w. multiview and entities" surpasses "w. multiview", further confirming the effectiveness of integrating retrieved semantic entities as textual prompts. Notably, the models with retrieved regional entities (i.e., textual prompts) demonstrate better performance than the models with only visual prompts.

TABLE IV

ABLATION STUDIES ON MIMIC-CXR. MULTIVIEW MEANS THE MULTIVIEW INTEGRATION. TEMP. MEANS TEMPORAL INCORPORATION. ENTITY MEANS THE INTEGRATION OF SEMANTIC RETRIEVAL. BL-N, MTR, RL, PRE., REC., RATE., CHEX., RADG MEAN BLEU-N, METEOR, ROUGE-L, PRECISION, RECALL, RATESCORE, CHEXBERT VECTOR SIMILARITY AND RADGRAPH-F1 RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Multiview	Temp.	Entity	BL-1	BL-2	BL-3	BL-4	MTR	RL	Pre.	Rec.	F1	RaTE.	CheX.	RadG.
✗	✗	✗	0.397	0.250	0.172	0.125	0.155	0.283	0.476	0.322	0.384	0.495	0.373	0.202
✓	✗	✗	0.412	0.263	0.181	0.131	0.163	**0.292**	0.522	0.405	0.456	0.507	0.390	0.209
✗	✓	✗	0.411	0.260	0.178	0.128	0.162	0.285	0.525	0.388	0.446	0.510	0.383	0.211
✗	✗	✓	0.406	0.262	0.183	0.134	0.161	0.291	0.530	0.392	0.451	0.509	0.386	0.216
✓	✓	✗	0.405	0.259	0.180	0.131	0.162	0.291	0.519	0.397	0.450	0.511	0.387	0.214
✓	✗	✓	0.419	0.268	0.185	0.135	0.164	0.286	**0.536**	0.405	0.461	0.513	0.392	0.217
✓	✓	✓	**0.422**	**0.271**	**0.188**	**0.137**	**0.165**	0.290	0.527	**0.412**	**0.462**	**0.516**	**0.397**	**0.219**

TABLE V

ABLATION STUDIES ON IU X-RAY. "W." MEANS "WITH". B-N, MTR, RL, RT, CX, RG REPRESENT BLEU-N, METEOR, ROUGE-L, RATESCORE, CHEXBERT VECTOR SIMILARITY AND RADGRAPH-F1 RESPECTIVELY. THE BEST RESULTS ARE MARKED IN BOLD

Method	B-3	B-4	MTR	RL	RT	CX	RG
baseline	0.229	0.173	0.206	0.374	0.634	0.599	0.302
w. multiview	0.230	0.175	0.205	0.382	0.648	0.605	0.333
w. entity	0.238	0.178	0.213	0.377	0.652	0.608	0.330
STREAM	**0.248**	**0.188**	**0.215**	**0.387**	0.659	0.615	0.357

TABLE VI

RESULTS CONCERNING LLM FINE-TUNING. * INDICATES THAT THE DECODER WAS FINE-TUNED USING LORA. B-N, MTR, RL, RT, CX, RG DENOTE BLEU-N, METEOR, ROUGE-L, RATESCORE, CHEXBERT VECTOR SIMILARITY AND RADGRAPH-F1 RESPECTIVELY

Dataset	B-3	B-4	MTR	RL	RT	CX	RG
MIMIC-CXR*	0.174	0.125	0.163	0.288	0.504	0.375	0.199
MIMIC-CXR	0.188	0.137	0.165	0.290	0.516	0.397	0.219
IU X-Ray*	0.245	0.184	0.203	0.369	0.650	0.603	0.321
IU X-Ray	0.248	0.188	0.215	0.387	0.659	0.615	0.357

We can draw a conclusion that the entities encapsulated with chest anatomical knowledge do help the fine-grained regional understanding.

In Table V, the performance improvement concerning the NLG metrics, such as BLEU-4, and METEOR, is modest. However, more improvements are observed in the performance concerning RaTEScore, CheXbert vector similarity, and RadGraph-F1. This could be attributed to the fact that the IU X-Ray dataset contains a larger proportion of normal CXR images compared with MIMIC-CXR, and it is hard for the NLG metrics to distinguish between clinical semantic differences in generated text. On the other hand, the radiology-specific metrics, which focus on the clinical and anatomical accuracy of the generated reports, show clearer improvements, suggesting the model's ability to capture detailed clinical entities and relationships. Overall, the components play an important role in enhancing the model's ability to generate more accurate and contextually rich CXR reports, with the best performance achieved when all components are employed.

Additionally, we conduct experiments to evaluate the performance of our STREAM under two conditions: with the decoder frozen and with the decoder set to be trainable (fine-tuned by LoRA). The results, presented in Table VI, indicate that the fine-tuned decoder performs worse than our STREAM with a frozen LLM decoder. This may be because cross-modal training compromises the language capabilities of the LLM.

E. Case Studies

In Fig. 5, we present an example of reports generated by STREAM and the ablation models, including "Baseline", "w. Multiview" and "w. Spatio-temporal Dynamics" (i.e., multiview spatio-temporal fusion). As can be seen, STREAM generates abnormalities accurately covering almost all diseases, except the "atelectasis". In contrast, the baseline correctly identifies "pacemaker" and "cardiomegaly", but incorrectly includes the diseases "vascular congestion" and "degenerative changes" while missing the descriptions for "calcified aorta", "hyperinflated lungs" and "atelectasis". In "w. Multiview", "pacemaker", "hyperinflated lungs" and "cardiomegaly" are correctly generated, while "calcified aorta" and "atelectasis" are missed. In "w. Spatio-temporal Dynamics", "pacemaker", "cardiomegaly" and "calcified aorta" are correctly generated, but "hyperinflated lungs" and "atelectasis" are missed. Furthermore, except for the accuracy analysis, we analyze the multi-view and temporal dynamic information. In Fig. 5, we can observe that STREAM and "w. Spatio-temporal Dynamics" generate the temporal dynamic descriptions accurately, and "w. Multiview" and "w. Spatio-temporal Dynamics" generate multiview information accurately, such as "PA and lateral views of the chest provided". Overall, our STREAM generates the report with rich disease information and temporal progressions. Also, we can observe that the retrieved entities contain accurate and essential anatomical disease information, which could further augment the generation process.

In Fig. 6, we showcase reports generated by R2Gen [5], R2GenGPT [23], PromptMRG [17], and our STREAM. The figure presents multiview images from the current study. Except our STREAM, the other comparison methods take only a single image for generation. As shown in Fig. 6 (a) and (b), STREAM accurately captures view-specific information, such as "AP upright and lateral views of the chest provided". Furthermore, STREAM excels at generating precise temporal details. In Fig. 6 (a), for instance, STREAM includes temporal information akin to those in the original reports, such as "There is a small right-sided pleural effusion which has increased in size since the prior examination" and "Moderate cardiomegaly is unchanged". In terms of disease captioning, our STREAM demonstrates superior accuracy compared with

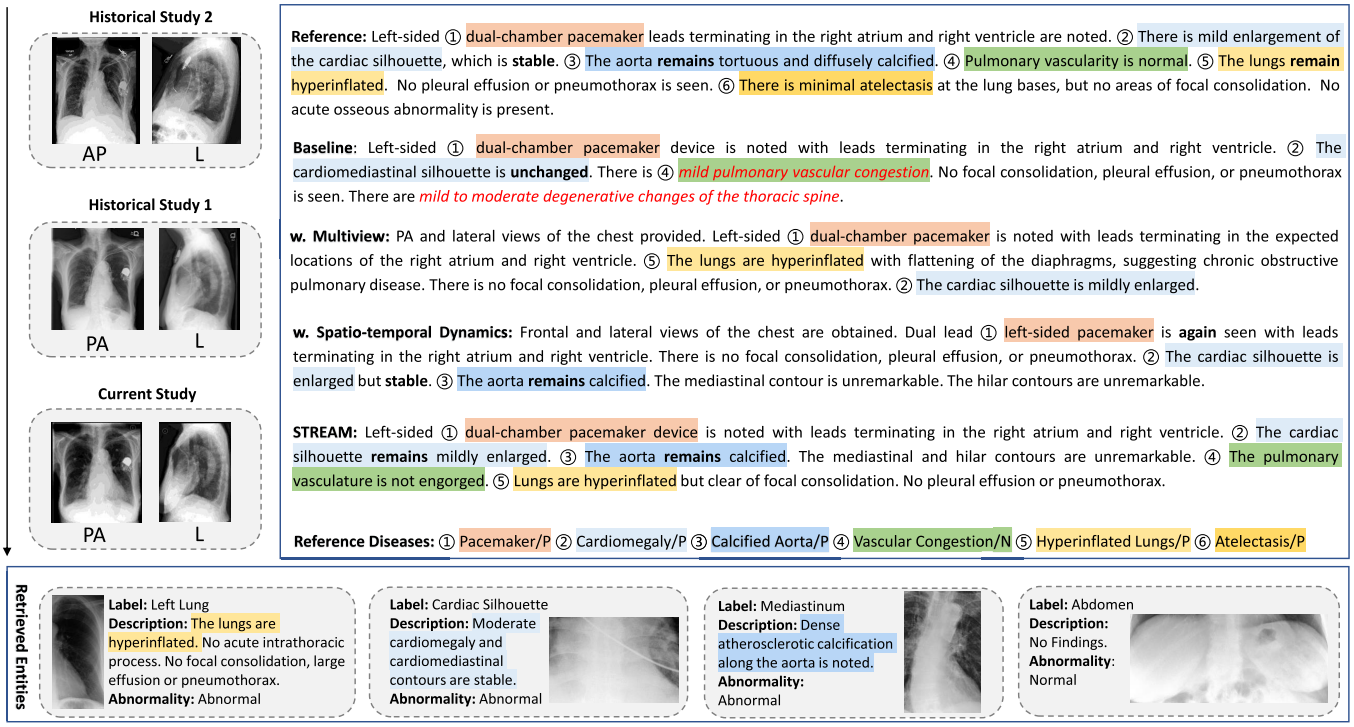

Fig. 5. The case study of our STREAM. The input image sequences are shown on the left. The reference report and generated reports are presented on the right. The retrieved regional semantic entities are presented at the bottom. Descriptions of a specific disease are highlighted in the same color. For each disease, "P" means a positive finding and "N" means a negative finding. "w." means "with". Temporal dynamics in the reports are highlighted in bold. Incorrectly generated disease descriptions are marked in red.

R2Gen that produces more incorrect disease captions. Additionally, when compared with R2GenGPT and PromptMRG, our STREAM generates fewer false positive disease captions.

In Fig. 7, we analyse the impact of multiview integration (i.e., how different views contribute to the report generation). In this case, images from a single study are shown in the top left, with the reference report, generated reports displayed on the right and the retrieved entities on the bottom. We performed the case study using three variants: a single frontal view, multi-view integration, and multi-view integration combined with retrieved entities for report generation. As shown, in "Single Frontal View", the model accurately identifies "Cardiomegaly" but fails to detect "small pleural effusion". In "Multiview Integration" and "Multiview and Retrieval Integration" scenarios, the model successfully generates both "Cardiomegaly" and "small pleural effusion". Notably, in this case as shown in Fig. 7, "small pleural effusion" is difficult to diagnose from the frontal view alone. The retrieved entities from the frontal view show no evidence of pleural effusion in the lungs. However, by integrating the lateral view, our method can generate the description "Blunting of the posterior costophrenic angles on the lateral view may be due to trace pleural effusions" accurately, indicating that multiview integration plays a crucial role in accurate disease diagnosis. By incorporating multiple perspectives, the clinicians as well as AI models could gain a comprehensive understanding of anatomical location (such as the location of a lung nodule) and detect abnormalities that may not be visible in a single view (such as small pleural effusions).

TABLE VII
HUMAN EVALUATIONS ON REPORTS GENERATED BY DIFFERENT METHODS. VALUES ARE IN PERCENTAGE %. ("ST" MEANS STREAM, R2GPT REFERS TO R2GENGPT, MRG REFERS TO PROMPTMRG)

Metrics	ST vs. R2Gen			ST vs. R2GPT			ST vs. MRG		
	loss	tie	win	loss	tie	win	loss	tie	win
Correctness	27.7	25.0	47.3	36.0	19.7	44.3	38.7	19.7	42.7
Coverage	28.7	19.7	51.7	41.3	12.3	46.3	45.0	15.3	39.7

F. Human Evaluation

In this section, we conduct human evaluations to assess the efficacy of STREAM. We randomly select 100 samples from the testing set of MIMIC-CXR for evaluation. Three medical doctors are invited to compare reports generated by STREAM with those from "R2Gen", "PromptMRG" and "R2GenGPT". Each comparison includes the report from STREAM, the report from the comparison method, the reference report, and the corresponding images. Doctors rate each comparison as a "win", "loss" or "tie" based on correctness and coverage. A "win" indicates better alignment with the reference report and image in terms of correctness (accuracy of normal and abnormal findings) or coverage (the rate of abnormalities generated compared to the reference). A "tie" is rated when reports show similar correctness or coverage. We then compute average percentages for "win", "tie" and "loss" outcomes. Results are presented in Table VII. Compared with R2Gen and R2GenGPT, our STREAM achieves better performance in both correctness and coverage. Compared with PromptMRG, our STREAM achieves better performance in correctness,

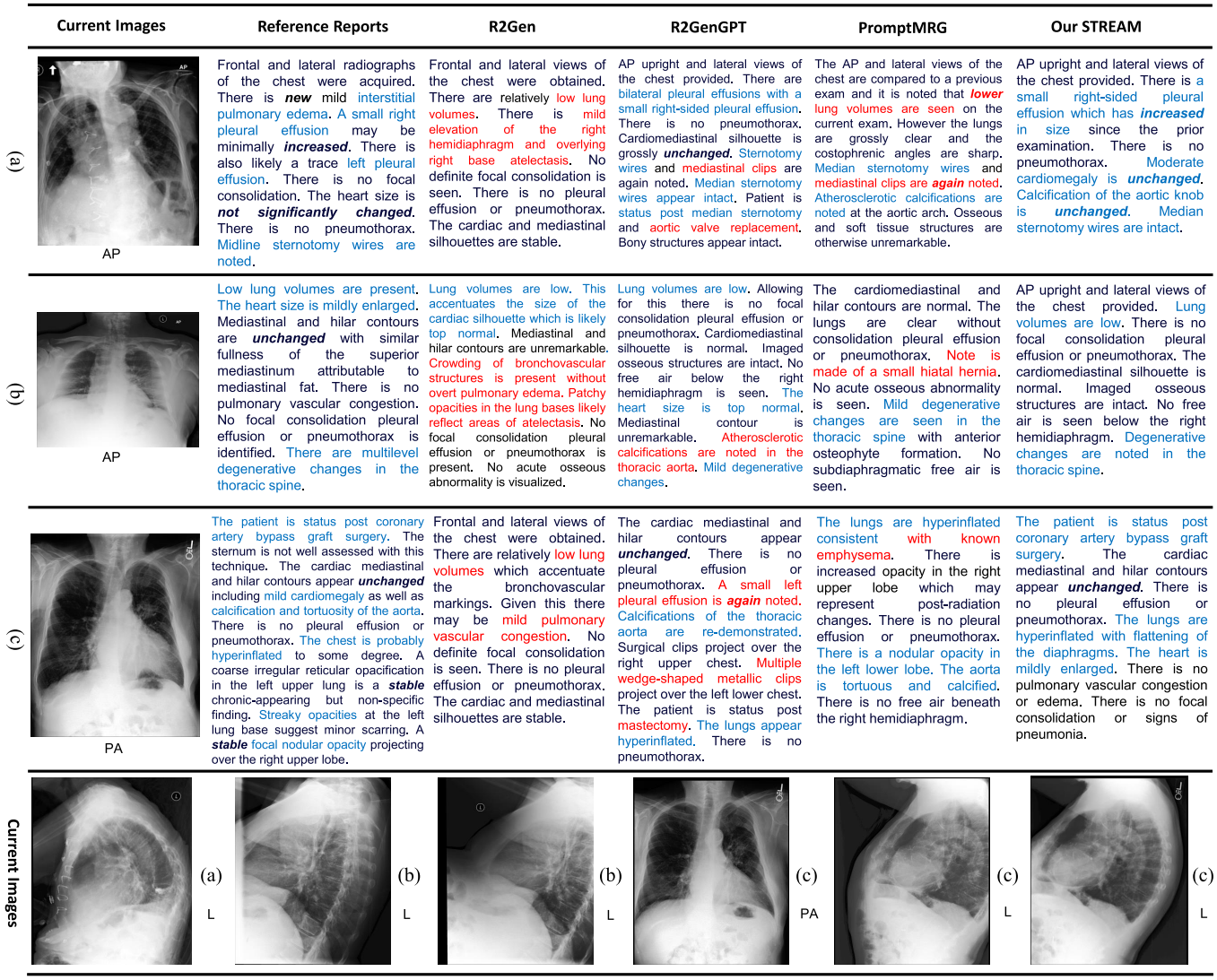

Fig. 6. Examples of CXR reports generated by different methods, i.e., R2Gen [5], R2GenGPT [23], PromptMRG [17] and our STREAM. AP, PA, and L mean the anteroposterior, posteroanterior and lateral views. Blue words are correct disease descriptions, while red words are incorrect disease descriptions. The normal findings are not highlighted. Italicized and bold words show the temporal information. In the first case, the reference report overlooks the calcification of the aorta, while our STREAM and PromptMRG generate the calcification accurately. Please zoom in for details.

TABLE VIII

EXPERIMENTS ON USING VARYING NUMBERS OF HISTORICAL STUDIES FOR REPORT GENERATION. "NUMBER OF HS" MEANS NUMBERS OF HISTORICAL STUDIES. BL-N, MTR, RL, PRE., REC., RATE., CHEX., RADG REPRESENT BLEU-N, METEOR, ROUGE-L, PRECISION, RECALL, RATESCORE, CHEXBERT VECTOR SIMILARITY AND RADGRAPH-F1 RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Number of HS	BL-1	BL-2	BL-3	BL-4	MTR	RL	Pre.	Rec.	F1	RaTE.	CheX.	RadG.
0 study	0.419	0.268	0.185	0.135	0.164	0.286	0.536	0.405	0.461	0.513	0.392	0.217
1 study	0.411	0.265	0.185	0.136	0.164	0.296	0.526	0.409	0.460	0.509	0.389	**0.220**
2 studies	**0.422**	**0.271**	**0.188**	**0.137**	0.165	0.290	0.527	0.412	0.462	**0.516**	0.397	0.219
3 studies	0.416	0.268	0.186	0.136	**0.166**	**0.297**	**0.538**	**0.425**	**0.475**	0.515	**0.400**	0.219
4 studies	0.406	0.262	0.182	0.133	0.162	**0.297**	0.530	0.407	0.551	0.506	0.392	0.209
5 studies	0.401	0.256	0.178	0.129	0.159	0.291	0.506	0.350	0.414	0.502	0.379	0.207

while PromptMRG achieves better performance in terms of the coverage.

G. Other Experiments

We further conduct experiments to evaluate the model's effectiveness in CXR report generation using varying numbers of historical studies on MIMIC-CXR. Notably, the historical studies were selected based on the timeline of the photograph's examination, prioritizing studies temporally closer to the current one. Since different patients have different number of examinations, we have padded the visual input with the current study for cases with insufficient historical studies. Table VIII presents the results of experiments conducted to evaluate the impact of varying numbers of historical studies (HS) on CXR

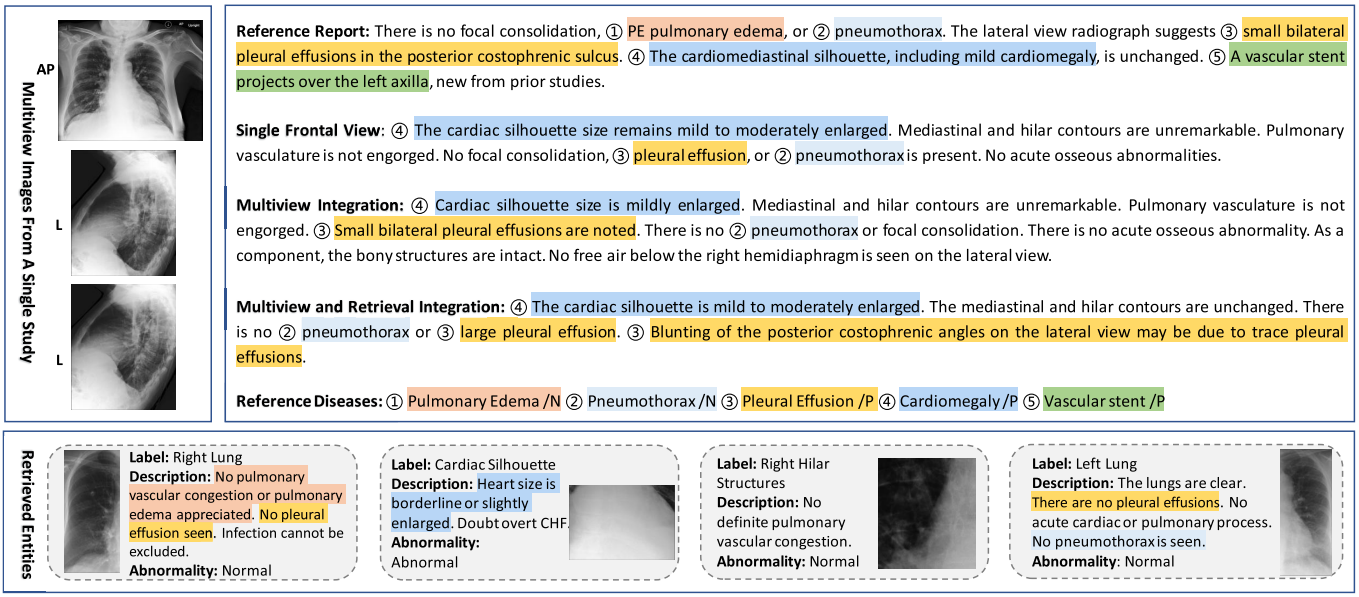

Fig. 7. The case study of our method focused on the multiview integration. Images from a single study are shown in the top left, with the reference report and generated reports displayed on the right. The retrieved regional semantic entities are presented at the bottom. Descriptions of a specific disease are highlighted in the same color. For each disease, "P" means a positive finding and "N" means a negative finding.

TABLE IX
EXPERIMENTS ON STREAM USING DIFFERENT METHODS FOR GENERATING SPATIALLY CONDENSED VISUAL TOKENS. BL-N, MTR, RL, PRE., REC., RATE., CHEX., RADG REPRESENT BLEU-N, METEOR, ROUGE-L, PRECISION, RECALL, RATESCORE, CHEXBERT VECTOR SIMILARITY AND RADGRAPH-F1 RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	BL-1	BL-2	BL-3	BL-4	MTR	RL	Pre.	Rec.	F1	RaTE.	CheX.	RadG.
Addition	0.419	0.265	0.181	0.130	0.162	0.287	0.492	0.377	0.427	0.497	0.376	0.203
Concatenation	0.416	0.263	0.181	0.131	0.163	0.287	0.478	0.414	0.444	0.495	0.379	0.209
Perceiver-Flow	**0.426**	0.269	0.185	0.134	**0.166**	0.289	0.501	**0.433**	**0.464**	0.507	0.383	0.213
Perceiver-All	0.418	0.264	0.182	0.132	0.162	0.288	0.509	0.389	0.441	0.499	0.386	0.212
Centered MHA	0.422	**0.271**	**0.188**	**0.137**	0.165	**0.290**	**0.527**	0.412	0.462	**0.516**	**0.397**	**0.219**

report generation. From Table VIII, we observe that using historical studies improves performance across most metrics. Specifically, utilizing 1, 2, or 3 historical studies leads to better results than not using any historical studies (0 studies) and also outperforms using 4 or 5 historical studies. For instance, the highest performance in BLEU scores and RaTEScore is achieved with 2 studies. The highest F1 score and RadGraph-F1 are achieved with 3 historical studies. Interestingly, while increasing the number of historical studies (e.g., 4 or 5) does not consistently improve the performance. This is because the report of the current study is written by radiologists according to clinical findings from latest studies, such as the description "a left internal jugular central venous catheter device has been removed in the interval". In routine practice, radiologists typically refer to 1-3 historical studies. More recent studies provide more relevant and up-to-date clinical context. It is also observed that the inference complexity increases as the number of historical studies increases. This suggests a trade-off between performance and computational efficiency when more historical studies are utilized.

Besides, we have conducted a series of experiments to explore the effectiveness of different methods for generating spatially condensed visual tokens, with the results presented in Table IX. Experiments are conducted on MIMIC-CXR. The aim of these experiments is to evaluate how various strategies for generating these tokens impact the performance of our STREAM framework. The methods include "Addition", "Concatenation", "Perceiver-All", "Perceiver-Flow", and our "Centered MHA". "Addition" means to directly add multiview image tokens encoded by a shared SwinTransformer [41]. "Concatenation" means to directly concatenate the multiview image tokens. "Perceiver-All" employs a set of learnable query tokens to conduct attention across the concatenated image tokens with residual connections. "Perceiver-Flow" initializes the process by using a set of learnable query tokens to conduct attention on the first set of image tokens, then utilizes the resulting tokens to conduct attention on the subsequent sets of image tokens in a sequential manner. It is noted that both "Perceiver-All" and "Perceiver-Flow" are based on the perceiver resampler proposed in the Flamingo [62]. Our "Centered MHA" is detailed in Fig. 3. Overall, our "Centered MHA" and the "Perceiver-Flow" methods demonstrate superior performance compared with the straightforward approaches of simply adding or concatenating image tokens. The "Perceiver-All" method performs worse than "Perceiver-Flow", suggesting that progressively integrating image tokens across different attention layers yields better results. It is also observed in our experiments that the residual connec-

tions in "Perceiver-All", "Perceiver-Flow" and Our "Centered MHA" are important. Without the residual connections, the performance drops sharply. An additional observation from our research is that there is a direct correlation between the number of tokens fed into the LLMs and the time of the inference process. Consequently, methods such as "Perceiver-All" and "Concatenation" which involve a larger number of tokens, result in longer inference time compared to "Perceiver-Flow" and "Centered MHA". This finding underscores the importance of token efficiency in the context of computation and suggests that more effective token compression strategies can lead to faster inference.

H. Limitations

In this section, we discuss the limitations of the knowledge bank construction and the token packer. 1) Since the Chest ImaGenome dataset is build upon only the frontal views. Therefore, the anatomical regions from the lateral views have not yet been included in our knowledge bank. 2) According to the Chest ImaGenome dataset, we select anatomical regions by eliminating overlapping and less significant regions. While chest X-rays contain numerous anatomical regions, this paper focuses on the most prominent ones: "right lung", "right hilar structures", "right apical zone", "left lung", "left hilar structures", "left apical zone", "trachea", "spine", "mediastinum", "cardiac silhouette", and "abdomen". In the future, we plan to expand the knowledge bank to include additional anatomical regions, such as "clavicle" and "hemidiaphragms". 3) In this paper, we capture the spatio-temporal visual dynamics using centered multi-head attention, concatenation, and addition operations. In the future, a more efficient token packer for spatio-temporal integration could be designed to enhance inference efficiency. Additionally, the integration of retrieved entities could be optimized to further improve inference speed, as lengthy entity descriptions introduce inefficiencies for the LLM decoder during inference.

V. Conclusion

In this paper, we present STREAM, a novel approach for chest X-ray report generation that integrates multi-view spatio-temporal information and cross-modal anatomical knowledge to mimic the clinical diagnosis. Our method integrates both current and historical studies to provide a comprehensive interpretation of the present condition, utilizing a large language model (LLM) as the decoder. We capture spatio-temporal visual dynamics via a token packer as visual prompts, and retrieve region-level semantic entities from our pre-established knowledge bank as textual prompts. The knowledge bank is constructed to encapsulate anatomical region knowledge of chest X-rays into structured semantic entities. By integrating spatio-temporal cues and anatomical knowledge into the generation, our method not only improves the accuracy but also aligns well with practical workflows. In our future works, we will explore more effective algorithms for the token compression of spatio-temporal CXR data. Additionally, we will consider adding reasoning strategies into the generation process, such as the chain-of-though.

References

[1] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7433–7442.

[2] B. Yan and M. Pei, "Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2982–2990.

[3] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11558–11567.

[4] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3334–3343.

[5] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1439–1449.

[6] A. E. W. Johnson et al., "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *Sci. Data*, vol. 6, p. 317, Jul. 2019.

[7] Q. Zhu, T. S. Mathai, P. Mukherjee, Y. Peng, R. M. Summers, and Z. Lu, "Utilizing longitudinal chest X-rays and reports to pre-fill radiology reports," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, vol. 14224, Jan. 2023, pp. 189–198.

[8] W. Hou, Y. Cheng, K. Xu, W. Li, and J. Liu, "RECAP: Towards precise radiology report generation via dynamic disease progression reasoning," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 2134–2147.

[9] Z. Liu, Z. Zhu, S. Zheng, Y. Zhao, K. He, and Y. Zhao, "From observation to concept: A flexible multi-view paradigm for medical report generation," *IEEE Trans. Multimedia*, vol. 26, pp. 5987–5995, 2024.

[10] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5904–5914.

[11] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 448–458.

[12] H. Shen, M. Pei, J. Liu, and Z. Tian, "Automatic radiology reports generation via memory alignment network," in *Proc. 38th AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 4776–4783.

[13] K. Zhang et al., "Semi-supervised medical report generation via graph-guided hybrid feature consistency," *IEEE Trans. Multimedia*, vol. 26, pp. 904–915, 2024.

[14] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "ORGAN: Observation-guided radiology report generation via tree reasoning," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1–17.

[15] Y. Yang et al., "Token-mixer: Bind image and text in one embedding space for medical image reporting," *IEEE Trans. Med. Imag.*, vol. 43, no. 11, pp. 4017–4028, Nov. 2024.

[16] C. Lyu et al., "Automatic medical report generation combining contrastive learning and feature difference," *Knowl.-Based Syst.*, vol. 305, Dec. 2024, Art. no. 112630.

[17] H. Jin, H. Che, Y. Lin, and H. Chen, "PromptMRG: Diagnosis-driven prompts for medical report generation," in *Proc. 38th AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 2607–2615.

[18] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102510.

[19] K. Zhang et al., "Attribute prototype-guided iterative scene graph for explainable radiology report generation," *IEEE Trans. Med. Imag.*, vol. 43, no. 12, pp. 4470–4482, Dec. 2024.

[20] L. Xu, B. Liu, A. H. Khan, L. Fan, and X.-M. Wu, "Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark," 2023, *arXiv:2306.06494*.

[21] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[22] Z. Chen et al., "CheXagent: Towards a foundation model for chest X-ray interpretation," 2024, *arXiv:2401.12208*.

[23] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," *Meta-Radiol.*, vol. 1, no. 3, Nov. 2023, Art. no. 100033.

[24] S. L. Hyland et al., "MAIRA-1: A specialised large multimodal model for radiology report generation," 2023, *arXiv:2311.13668*.

[25] S. Lee, W. J. Kim, and J. C. Ye, "LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation," in *Proc. 12th Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–11.
[26] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2023, pp. 543–553.
[27] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, vol. 202, Jan. 2023, pp. 19730–19742.
[28] Q. Zhu et al., "Spatio-temporal graph hubness propagation model for dynamic brain network classification," *IEEE Trans. Med. Imag.*, vol. 43, no. 6, pp. 2381–2394, Jun. 2024.
[29] N. Ahmadi, M. Y. Tsang, A. N. Gu, T. S. M. Tsang, and P. Abolmaesumi, "Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 366–376, Jan. 2024.
[30] Y. Wang, Z. Ye, M. Wen, H. Liang, and X. Zhang, "TransVFS: A spatio-temporal local–global transformer for vision-based force sensing during ultrasound-guided prostate biopsy," *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103130.
[31] S. Bannur et al., "Learning to exploit temporal structure for biomedical vision-language processing," 2023, *arXiv:2301.04558*.
[32] F. Wang, S. X. Du, and L. Yu, "HERGen: Elevating radiology report generation with longitudinal data," in *Proc. 18th Eur. Conf. Comput. Vis.-ECCV*, Nov. 2024, pp. 183–200.
[33] X. Mei, R. Mao, X. Cai, L. Yang, and E. Cambria, "Medical report generation via multimodal spatio-temporal fusion," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 4699–4708.
[34] P. Zhao et al., "Retrieval-augmented generation for AI-generated content: A survey," 2024, *arXiv:2402.19473*.
[35] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 9802–9822.
[36] N. Carlini et al., "Extracting training data from large language models," 2020, *arXiv:2012.07805*.
[37] Y. Yan and W. Xie, "EchoSight: Advancing visual-language models with Wiki knowledge," 2024, *arXiv:2407.12735*.
[38] S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Retrieval-augmented transformer for image captioning," 2022, *arXiv:2207.13162*.
[39] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "Smallcap: Lightweight image captioning prompted with retrieval augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2840–2849.
[40] M. Kong, Z. Huang, K. Kuang, Q. Zhu, and F. Wu, "TranSQ: Transformer-based semantic query for medical report generation," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, vol. 13438. Cham, Switzerland: Springer, Jan. 2022, pp. 610–620.
[41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[43] J. T. Wu et al., "Chest ImaGenome dataset for clinical reasoning," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, Jan. 2021, pp. 1–11.
[44] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3876–3887.
[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[46] M. Douze et al., "The faiss library," 2024, *arXiv:2401.08281*.
[47] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.
[48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[49] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[50] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization ACL*, 2004, pp. 74–81.
[51] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, Jun. 2007, pp. 228–231.
[52] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "RaTEScore: A metric for radiology report generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2024, pp. 15004–15019.
[53] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, and M. Lungren, "Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1500–1519.
[54] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, Jan. 2021, pp. 1–12.
[55] F. Yu et al., "Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, Sep. 2023, Art. no. 100802.
[56] P. Zhang, G. Zeng, T. Wang, and W. Lu, "TinyLlama: An open-source small language model," 2024, *arXiv:2401.02385*.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.
[58] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–7.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[60] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[61] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–8.
[62] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," 2022, *arXiv:2204.14198*.