TREAM: A Temporal Reasoning and Evolutionary Analysis Model for Radiology Report Generation

Zhen Zhang[1,2,3], Zhaoyang Wang[1,2,3], Jing Zhao[1,2,3], Song Liu[1,2,3*]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology,
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[3]Shandong Provincial Key Laboratory of Industrial Network and Information System Security,
Shandong Fundamental Research Center for Computer Science, Jinan, China
Email:{10431240225,10431250082}@stu.qlu.edu.cn, zjstudent@126.com, liusong@qlu.edu.cn

***Abstract*—Radiology report generation requires advanced medical image analysis, effective temporal reasoning, and accurate use of radiological terminology. Although multimodal large language models (MLLMs) align with pre-trained vision encoders to enhance visual–language understanding, most existing methods rely on single-image analysis to process images, failing to fully leverage temporal information in multimodal medical datasets. In addition, insufficient cross-modal alignment can lead to imprecise use of radiological terminology, and the lack of intermediate diagnostic reasoning and incomplete utilization of reference report content further impair interpretability, potentially resulting in hallucinations. These limitations ultimately prevent the models from generating radiology reports that are accurate in clinical terminology and capable of capturing disease progression. To address these limitations, we propose TREAM (Temporal Reasoning and Evolutionary Analysis Model), a temporal-aware MLLM that explicitly captures disease progression from longitudinal imaging features, enhances image-text alignment through an additional contrastive learning loss, and employs a five-step Chain-of-Medical-Thought to generate reports, thereby improving interpretability and reducing hallucinations. We experimented and evaluated our model on the MIMIC-CXR dataset, and the results demonstrate that TREAM achieves superior performance across both natural language generation and clinical effectiveness metrics, highlighting its applicability to longitudinal radiology report generation.**

***Index Terms*—Radiology Report Generation, Temporal Analysis, Multimodal Large Language Models, Contrastive Learning, Chain-of-Medical-Thought**

I. Introduction

Radiology Report Generation is a key task in medical image analysis, with the aim of automatically generating complete and clinically accurate diagnostic reports to help radiologists improve work efficiency and diagnostic consistency [1], [2]. The core challenge of this task lies in transforming high-dimensional image features from medical images into textual descriptions that conform to clinical standards. This requires models not only to extract accurate visual features, but also to possess clinical knowledge to interpret pathological findings, as well as the ability to organize and generate logically coherent and diagnostically precise reports. However, the rapid growth of medical imaging data, combined with a shortage of qualified radiologists, has driven the demand for automatic radiology report generation [3].

Deep learning has been widely applied to radiology report generation, ranging from CNN–RNN frameworks [4] to the integration of advanced Transformer architectures [5]. Nevertheless, these methods show limited effectiveness in clinical practice. With the development of large language models, recent research has explored the use of MLLMs in radiology report generation. KARGEN [6] uses a medical knowledge graph with frozen large language models (LLMs) to extract disease features from chest radiographs for more accurate report generation. Med-Gemini [7] fine-tunes LLMs on diverse medical data to perform report generation and diagnostics. With their larger capacity, enhanced generalization, and improved multimodal integration, these models have shown some improvement in radiology report generation.

Despite recent advances in multimodal models, radiology report generation involves unique clinical requirements. In particular, models must be able to integrate information from longitudinal imaging studies, maintain precise medical terminology, and produce coherent diagnostic reasoning that accurately reflects disease progression over time. Existing approaches do not fully satisfy these requirements [8].

Although recent methods have made progress in radiology report generation, existing approaches still have the following problems: 1) Previous research relies mainly on final-layer encoder features and uses ad hoc temporal fusion strategies, which ignore correlations between different images and lack consistent temporal modeling [9]. 2) Existing models have limited understanding of medical terminology and patholog-

This work was supported by the Key Research and Development Program (Major Innovation Projects) in Shandong Province under Grant 2021CXGC010102, and the Key Research and Development Program in Shandong Province (The Project for Enhancing the Innovation Capacity of Technology-based Small and Medium-sized Enterprises) under Grant 2025TSGCCZZB0124.

*Corresponding author: liusong@qlu.edu.cn

ical concepts, which may lead to a semantic gap between visual features and medical language. 3) Existing end-to-end approaches directly map images to reports without explicit reasoning, which can result in outputs that are less interpretable and prone to hallucinations [10].

To address these challenges, we propose **TREAM**, a temporal-aware MLLM for radiology report generation. The main contributions are as follows:

- To capture temporal dependencies across multi-temporal images and disease evolution over time, we design a Temporal Aggregator module that includes Layer Feature Extraction, Dual-weighted Temporal Fusion, and Group Causal Transformer. This module can effectively exploit multi-layer visual representations and guide the model to focus on the most relevant historical information while preserving temporal causality across longitudinal exams.
- To ensure accurate medical terminology and strengthen cross-modal semantic consistency, we propose a Contrastive Learning Loss to align visual and textual representations, ensuring precise medical terminology and enabling the model to generate clinically accurate, semantically consistent reports.
- To improve diagnostic reasoning and reduce hallucinations, we propose a Chain-of-Medical-Thought consisting of modality recognition, anatomical structure identification, size and morphology assessment, abnormality detection, and comprehensive report generation, allowing the model to learn diagnostic reasoning that simulates the clinical thinking of radiologists.

II. RELATED WORK

Recent advances in MLLMs have demonstrated versatility in many tasks, including healthcare applications such as medical exams and radiology report generation. RaDialog [11] supports interactive report refinement for individual exams. R2GenGPT [12] achieves efficient fine-tuning by freezing language model parameters, while still focusing on the current image. CheX-agent [13] employs a multi-task learning framework, processing each radiograph independently. RadFM [14] fine-tunes a general image-text model on radiology-specific images, but does not take advantage of the longitudinal context. LLaVA-Rad [15] introduces a lightweight multimodal architecture with fine-grained labeling and automated evaluation, yet remains constrained to single image inputs. Despite these advances, these approaches are limited in their ability to utilize historical imaging, which restricts their applicability to longitudinal radiology report generation.

To address the challenges of longitudinal radiology report generation, which involves using historical chest X-rays and their reports to capture disease progression, HERGen [16] proposes a hierarchical encoding framework that learns multi-level temporal representations from sequential chest X-rays to model disease progression across exams, while Libra [17] introduces a priori perception mechanisms that integrate historical image context with current scans to enhance diagnostic accuracy. Although these methods have made some progress, they have not thoroughly explored the adaptation of LLMs to longitudinal medical data, missing the complex progression of diseases.

To improve radiology report generation in medical terminology and diagnostic reasoning, DCL [18] uses contrastive learning to align cross-modal features with textual semantics. However, it relies on general encoders, which limit the precise understanding of radiology-specific terminology. Methods such as M4CXR [19] improve interpretability through handcrafted inference steps, but do not leverage report content to guide diagnostic reasoning.. DAMPER [20] enhances predictions using a static medical knowledge graph, but lacks precise image-text alignment and adaptive diagnostic reasoning. They do not fully exploit radiology specific knowledge or ensure precise image-text alignment. Most models also bypass the intermediate diagnostic steps of radiologists, increasing the risk of factual and logical errors.

III. METHOD

Fig. 1 shows the overall architecture of TREAM, which includes image encoder, text encoder, Temporal Aggregator, Total Loss including Standard Cross-entropy Loss and Contrastive Learning Loss, and the pre-trained LLM for report generation. The image encoder processes the current chest radiograph together with multiple prior images, and the Temporal Aggregator integrates these hierarchical visual features into a temporally fused representation.

The aggregated features are then aligned with report text through a contrastive learning objective, while a five-step Chain-of-Medical-Thought is applied during training to provide structured supervision for diagnostic reasoning. The aligned features are finally fed into the pre-trained LLM to generate the radiology report.

A. Temporal Aggregator

The analysis of longitudinal chest radiographs requires capturing temporal evolution and relationships across examinations. To capture longitudinal changes across current and prior chest radiographs, the *Temporal Aggregator* bridges the image encoder and subsequent modules. It processes radiographs from the current exam and K prior time points to produce a unified representation that encodes temporally fused dependencies. As shown in Fig. 1, the *Temporal Aggregator* consists of three main components: the *Layer Feature Extraction*, which extracts multi-layer visual features from each image; the *Dual-weighted Temporal Fusion*, which adaptively integrates historical exams by considering both semantic similarity and temporal proximity; the *Group Causal Transformer*, which models longitudinal dependencies while maintaining temporal causality.

1) Layer Feature Extraction: Existing methods do not fully utilize multi-layer representations encoded by image encoders.To fully exploit these representations, we employ the RAD-DINO encoder [21], which consists of 12 hidden layers, and aggregate features from all layers. This captures detailed diagnostic information, such as the lesion boundaries

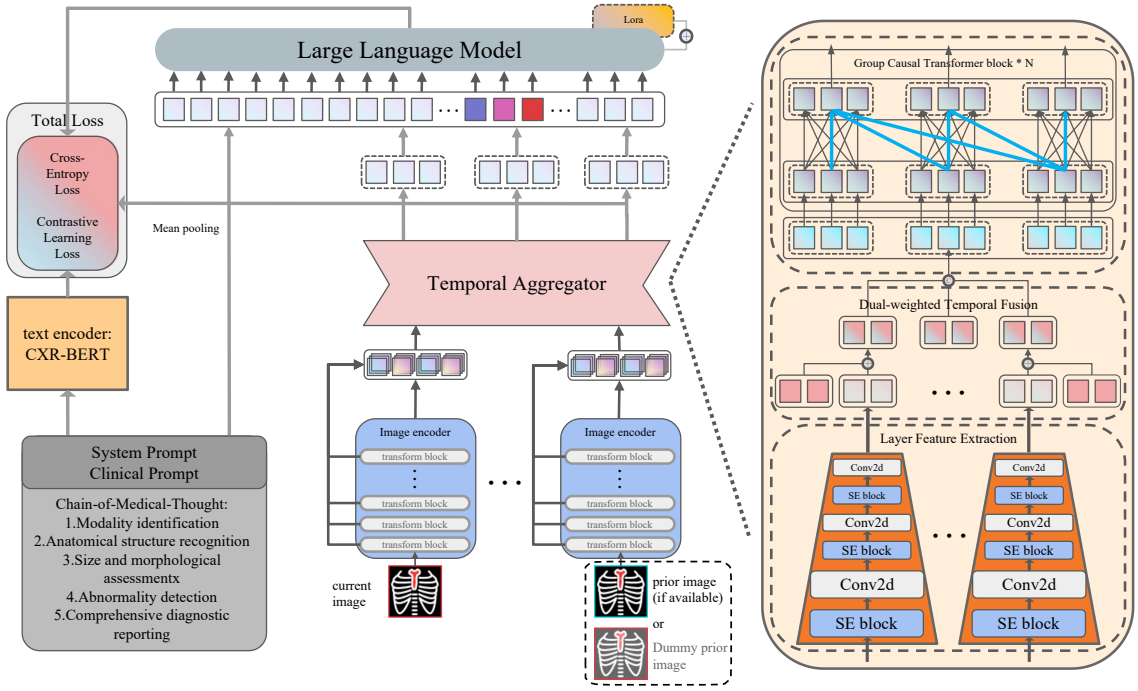


Fig. 1. The overall architecture of TREAM. Our model captures disease progression from longitudinal imaging features, enhances image-text alignment through contrastive learning objective, and employs a five-step Chain-of-Medical-Thought to generate reports.

and density variations, allowing efficient utilization of the information in the images.

Specifically, each input image is encoded in multi-layer features $F \in \mathbb{R}^{B\times H\times N\times D}$, where B is the batch size, $H = 12$ denotes the number of layers, N is the number of tokens per image, and D is the feature dimension.

Each stage applies a Squeeze-and-Excitation (SE) module for channel-wise feature recalibration. This module adaptively adjusts the importance of different feature channels to enhance meaningful signals, followed by GELU activation and a 1×1 convolution to align the feature dimensions. Following the compression path $12 \to 6 \to 3 \to 1$, the stages are:

$$F^{(1)} = \text{Conv2d}\left(\text{SE}_{12\to 6}(F)\right) \tag{1}$$

$$F^{(2)} = \text{Conv2d}\left(\text{SE}_{6\to 3}(F^{(1)})\right) \tag{2}$$

$$F^{(3)} = \text{Conv2d}\left(\text{SE}_{3\to 1}(F^{(2)})\right) \tag{3}$$

The final output is $F_{\text{out}} \in \mathbb{R}^{B\times N\times D}$. This design ensures that the output preserves the details of the image while maintaining the overall context.

2) Dual-weighted Temporal Fusion: The simple sequential processing method treats all historical images equally, which can obscure the progression of the disease. To address this, the *Dual-weighted Temporal Fusion* adaptively combines prior studies based on semantic similarity and temporal recency.

Specifically, F_{cur} represents the current feature and $F^{(i)}_{\text{prior}}$ the i-th historical feature ($i = 0$ corresponds to the earliest exam and $i = K - 1$ to the most recent). Cosine similarity measures how close each historical feature is to the current feature, and raises it to the eighth power acts as an empirical sharpening factor to emphasize highly relevant prior studies while suppressing weakly related ones:

$$\sigma_i = \left(\frac{\cos(F_{\text{cur}}, F^{(i)}_{\text{prior}}) + 1}{2}\right)^8 \tag{4}$$

where σ_i represents the similarity weight for the i-th historical study.To incorporate temporal information, a quadratic recency factor is applied:

$$r_i = \left(\frac{i+1}{K}\right)^2 \tag{5}$$

where r_i increases the weight for more recent studies.The final weight combines both factors along with a learnable prior bias b_{prior}:

$$w_i = \sigma_i \cdot r_i \cdot b_{\text{prior}} \tag{6}$$

Each historical feature is scaled accordingly:

$$F'^{(i)}_{\text{prior}} = w_i \cdot F^{(i)}_{\text{prior}} \tag{7}$$

This mechanism emphasizes prior images that are both similar to the current case and more recent, helping the model focus on the most relevant information.

3) Group Causal Transformer: Simple temporal fusion methods that allow each time point to attend to all others can leak future information, obscuring the true progression of the disease.

To model causal dependencies in medical imaging time series, we design the *Group Causal Transformer*, which limits information flow so that each time point can only use information from itself and previous time points, preventing future data from influencing the current prediction.

The adjusted historical features $\{F_{\text{prior}}^{(i)}\}_{i=0}^{K-1}$ and the current feature F_{cur} are stacked along the temporal dimension to form $F_{\text{temp}} = \text{stack}([F_{\text{prior}}^{(0)}, \ldots, F_{\text{prior}}^{(K-1)}, F_{\text{cur}}]) \in \mathbb{R}^{B\times T\times N\times D}$, where $T = K + 1$.

To help the model recognize temporal and spatial structure, F_{temp} is flattened from (B, T, N, D) to $(B, T \cdot N, D)$ by merging the temporal and spatial dimensions into a single sequence of tokens. Positional embeddings pos and temporal embeddings t are then added, producing the Transformer input: $X^{(0)} = \text{Flatten}(F_{\text{temp}}) + \text{pos} + t$.

The positional embeddings are learnable vectors that encode the position of each token in the flattened sequence, telling the model the relative order and spatial location of features. The temporal embeddings are learnable vectors for each time step, indicating from which time point each feature originates. Both types of embeddings are initialized as zero tensors and optimized during training.

The $X^{(0)}$ is then fed into the *Group Causal Transformer*, which uses multi-headed self-attention with a group causal mask $\mathbf{M} \in \{0, -\infty\}^{(TN)\times(TN)}$. This mask prevents the model from attending to future time steps and is inspired by the causal masking used in GPT. The attention weights are computed as follows:

$$\boldsymbol{\alpha}_{p,p'}^{(l,h)} = \text{Softmax}\left(\frac{\mathbf{q}_p^{(l,h)}(\mathbf{k}_{p'}^{(l,h)})^\top}{\sqrt{D_h}} + \mathbf{M}\right) \tag{8}$$

$$\mathbf{s}_p^{(l,h)} = \sum_{p'} \boldsymbol{\alpha}_{p,p'}^{(l,h)} \mathbf{v}_{p'}^{(l,h)} \tag{9}$$

where $\boldsymbol{\alpha}_{p,p'}^{(l,h)}$ indicates how much token p attends to token p' in head h and layer l, and $\mathbf{s}_p^{(l,h)}$ is the resulting weighted sum of value vectors.

For a given time index i, the valid attention region spans tokens from steps 0 to i, implemented by setting $\mathbf{M}[iN : (i+1)N, : (i+1)N] = 0$ and $-\infty$ elsewhere.

After multi-head attention, the concatenated features pass through an output projection, followed by a residual connection and a feed-forward network (FFN). Due to the short length of the radiological time series, only the $L = 2$ Transformer layers are used. The final output $X^{(L)} \in \mathbb{R}^{B\times(T\cdot N)\times D}$ represents the processed features after capturing temporal dependencies.

B. Contrastive Learning Loss

Previous research on radiology report generation is optimized mainly using standard cross-entropy loss $\mathcal{L}_{CE}$, which encourages the model to generate correct words by comparing predicted tokens with ground truth reports, but does not explicitly ensure alignment between medical images and their reports.

In our model, the temporally aggregated visual features $\mathbf{V} \in \mathbb{R}^{B\times N\times 768}$ serve two purposes: for report generation and for cross-modal alignment. Textual features from radiology reports serve as the reference for this alignment. In order to align the image information with the text of the radiology report, we design a contrastive learning objective inspired by InfoNCE.

In contrastive learning, visual features are projected onto a compact 128-dimensional shared space through a linear learning layer and averaged across the token dimension to obtain a global visual representation $\mathbf{D} \in \mathbb{R}^{128}$.This projection is used only for contrastive alignment, while the original visual features are retained for report generation. In parallel, the diagnostic report R is encoded by CXR-BERT, a medical language model pretrained in chest radiograph reports, which yields a textual representation $\mathbf{E} \in \mathbb{R}^{128}$. For a batch size of B , with embeddings $\{\mathbf{D}_r\}_{r=1}^{B}$ and $\{\mathbf{E}_r\}_{r=1}^{B}$, the contrastive learning loss is the following:

$$\mathcal{L}_{\text{Cont}} = \frac{1}{2B}\sum_{r=1}^{B}\left[-\log\frac{\exp\left(\text{sim}(\mathbf{D}_r, \mathbf{E}_r)/\tau\right)}{\sum_{s=1}^{B}\exp\left(\text{sim}(\mathbf{D}_r, \mathbf{E}_s)/\tau\right)} - \log\frac{\exp\left(\text{sim}(\mathbf{E}_r, \mathbf{D}_r)/\tau\right)}{\sum_{s=1}^{B}\exp\left(\text{sim}(\mathbf{E}_r, \mathbf{D}_s)/\tau\right)}\right] \tag{10}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. This formula measures how well each image-text pair matches compared to all other pairs in the batch.

The overall learning of the model is performed by minimizing a total loss that jointly optimizes report generation and image-text alignment. This total loss combines the standard cross-entropy loss $\mathcal{L}_{CE}$ and contrastive learning loss $\mathcal{L}_{\text{Cont}}$:

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{\text{Cont}} \tag{11}$$

where λ balances the importance of the contrastive loss relative to cross-entropy.

C. Chain-of-Medical-Thought

To improve the factual consistency between generated reports and imaging findings, we design a structured five-step *Chain-of-Medical-Thought* to allow our MLLM to simulate the clinical thinking of radiologists. This decomposes the diagnostic process into clinically grounded steps that closely follow standard radiological practice.

The *Chain-of-Medical-Thought* is applied only during training. It guides the model to learn and leverage the textual conventions of the ground truth reports, structuring the reasoning steps to correspond with the report content, and providing clear supervision for diagnostic reasoning.

Our *Chain-of-Medical-Thought* consists of five sequential steps:

(1) Modality identification. The imaging modality (e.g., X-ray or CT) is first identified to establish the diagnostic context.

(2) Anatomical structure recognition. The thoracic anatomical structures are identified using a predefined keyword database covering 19 entities. Relevant terms are extracted from reports through regular expression matching and normalized to ensure consistency.

(3) Size and morphological assessment. Organ size, shape, and visual clarity are assessed using five predefined categories: *normal*, *increased*, *decreased*, *clear*, and *fuzzy*. The corresponding descriptors are extracted and standardized via keyword matching.

(4) Abnormality detection. The pathological results are determined through a comprehensive database of 10 common chest diseases, each of which has multiple corresponding synonyms. To distinguish between positive and negative findings, the negation detector examines the three words preceding each pathological keyword (e.g., "without effusion" versus "with effusion"). In addition, temporal keywords such as "stable", "new", and "increased" are detected to capture disease progression, complementing the temporal modeling module.

(5) Comprehensive diagnostic reporting. Integrate the information extracted in the previous steps to generate a complete diagnostic report. The complete ground truth report is used as the training target to preserve clinical completeness.

For each report, entities are extracted through regular expression–based matching, and question–answer pairs are constructed using negation and temporal analysis. Our *Chain-of-Medical-Thought* ensures a strict correspondence between the reasoning chain and the ground truth report content, providing a clear supervisory mechanism for structured diagnostic reasoning.

IV. EXPERIMENT

A. Dataset Description

We train and evaluate our model on the MIMIC-CXR dataset [23], a large-scale chest X-ray report dataset containing 377,110 images and 227,827 radiology reports from 65,379 patients. Each radiology report typically consists of two main sections: *Findings*, which describe detailed observations from the imaging study, and *Impression*, which provides a concise summary and diagnostic conclusion. Based on this structured organization and following the proposed *Chain-of-Medical-Thought* described in Section III, each sample is transformed into a set of structured question–answer pairs to support step-wise diagnostic reasoning.

Since our model leverages longitudinal imaging information, some patients have varying numbers of historical images, resulting in incomplete prior studies for certain samples. To ensure a consistent temporal input format, missing historical images are filled with a *dummy prior image* created by replicating the most recent available image.

B. Experimental Settings

To train the TREAM model effectively, we design a two-stage training strategy while keeping the image encoder and the pre-trained LLM frozen throughout. Specifically, we adopt RAD-DINO as the image encoder to extract image features and Meditron-7B [24] as the backbone large language model for report generation. Each input consists of a fixed number of images determined by the *temporal_block_size* parameter(set to 3 in our experiments), which includes the current image and a number of prior images. The AdamW optimizer is used for all training stages.

Stage 1 (Temporal Feature Alignment). Only the *Temporal Aggregator* is trained using both the *Findings* and *Impression* sections of radiology reports, while the CXR-BERT text encoder remains frozen. This stage enables the model to establish robust correlations between image features and report semantics. Training is performed for one epoch with a learning rate of 2×10^{-5}.The contrastive loss weight is set to 0.5 to balance feature alignment and report reconstruction, and the temperature τ is fixed at 0.07, following common practice in contrastive learning to control the sharpness of the similarity distribution.

Stage 2 (Downstream Task Fine-tuning). The pre-trained LLM is fine-tuned via LoRA adapters (rank $r = 128$, scaling factor $\alpha = 256$), while the backbone parameters remain frozen. We use a relatively high rank to provide sufficient adaptation capacity for temporally grounded and clinically structured report generation. Training is conducted for three epochs with a learning rate of 1×10^{-5}. The contrastive loss weight is reduced to 0.2 to focus more on report generation, while the temperature τ remains fixed.

During inference, we employ a beam search strategy with a beam size of 3 to generate the reports.

C. Evaluation Metrics

We assess the performance of our model using both natural language generation (NLG) metrics and clinical efficacy (CE) metrics. For NLG evaluation, we employ BLEU-n [25], METEOR [26] and ROUGE-L [27] to measure the textual quality of the generated reports. For the evaluation of clinical effectiveness, we compute the precision, recall, and F1 scores. Specifically, both generated and ground truth reports are converted into 14 thoracic disease labels using CheXbert [28], and CE metrics are calculated based on these labels.

D. Comparative Experiments

In order to validate the effectiveness of our model, TREAM was compared with several different categories of medical report generation models:

Models that focus on temporal dependencies:

• R2Gen [29] fuses image and textual features in an end-to-end Transformer framework to capture temporal relationships in sequential exams.

• R2GenCMN [30] introduces a cross-modal memory network to retain historical information and support detailed image–text correspondence.

• HERGen [16] incorporates longitudinal patient visits to model temporal consistency across multiple exams.

TABLE I
NLG AND CE PERFORMANCE RESULTS ON MIMIC-CXR. ‡DATA SOURCED FROM OTHER STUDIES [22]. "–" INDICATES MISSING DATA. BEST PERFORMANCE SHOWN IN **BOLD**, SECOND-BEST UNDERLINED.

Method	NLG				CE		
	BLEU-1	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F1
R2Gen	0.339	0.103	0.138	0.279	0.297	0.189	0.193
R2GenCMN	0.345	0.101	0.140	0.274	0.354	0.271	0.275
M2Transformer	0.352	0.096	0.128	0.263	0.239	0.173	0.173
CheXagent‡	0.169	0.047	–	0.215	–	–	–
XProNet	0.303	0.091	0.128	0.268	0.419	0.230	0.242
Med-PaLM‡	0.323	0.115	–	0.275	–	–	–
DCL	0.263	0.071	0.117	0.211	0.303	0.232	0.229
HERGen	0.395	0.122	0.156	0.285	0.415	0.301	0.317
LLaVA-Rad	0.381	0.123	0.148	0.273	0.412	0.298	0.344
TREAM	**0.398**	**0.140**	**0.169**	**0.297**	**0.439**	**0.323**	**0.373**

Models that emphasize cross-modal alignment and semantic consistency :

• XProNet [31] combines cross-modal prototype learning with multi-label contrastive loss to alleviate category imbalance and feature confusion.

• DCL [18] applies a double contrastive loss to align visual and textual representations, improving clinical consistency.

Models that enhance diagnostic reasoning or reduce hallucinations:

• M2Transformer [32] extends the Transformer with a meshed memory mechanism to better capture contextual dependencies and reduce hallucinations.

• Med-PaLM [33] adapts the PaLM language model to the medical domain through knowledge distillation and alignment techniques to reduce factual errors in generated reports.

• CheXagent [13] uses a four-stage training strategy to improve clinical fidelity in chest X-ray interpretation.

• LLaVA-Rad [15] adapts a lightweight multimodal architecture to radiological data, enhancing reasoning over multimodal input and improving clinical decision support.

As shown in Tables I, our method achieves improvements across most evaluation metrics. On the MIMIC-CXR dataset, TREAM outperforms HERGen by 1.8% and 5.6% on BLEU-4 and F1, respectively, and surpasses LLaVA-Rad by 1.4% on BLEU-4 and 2.9% on F1. In addition, our model improves METEOR and ROUGE-L over XProNet by 1.5% and 2.1%, respectively, and enhances Precision and Recall in clinical effectiveness metrics by 2.4% and 2.2% compared to DCL.

These results indicate that leveraging TREAM's key mechanisms across temporal, reasoning, and alignment dimensions substantially benefits radiology report generation. Temporal feature fusion preserves consistency in disease progression, while cross-modal alignment ensures semantic and clinical accuracy. By combining these mechanisms, the model employs step-wise diagnostic reasoning, enabling it to generate reports that are coherent and closely reflect clinical findings. In general, the observed improvements in both natural language generation and clinical effectiveness metrics demonstrate the robustness and practical value of our approach.

These consistent improvements across multiple metrics indicate robust and stable performance.

E. Ablation Studies

TABLE II
ABLATION STUDY (NLG) RESULTS ON MIMIC-CXR DATASET.

Method	NLG			
	BLEU-1	BLEU-4	METEOR	ROUGE-L
Base Model	0.381	0.123	0.148	0.273
w/o LFE	0.394	0.132	0.168	0.294
w/o DTF	0.393	0.133	0.168	0.295
w/o GCT	0.391	0.131	0.167	0.292
w/o CLL	0.389	0.130	0.166	0.291
w/o CoMT	0.397	0.134	0.169	0.295
TREAM	**0.398**	**0.140**	**0.169**	**0.297**

TABLE III
ABLATION STUDY (CE) RESULTS ON MIMIC-CXR DATASET.

Method	CE		
	Precision	Recall	F1
Base Model	0.412	0.298	0.344
w/o LFE	0.439	0.320	0.371
w/o DTF	0.440	0.321	0.372
w/o GCT	0.432	0.318	0.367
w/o CLL	0.438	0.319	0.370
w/o CoMT	0.436	0.321	0.370
TREAM	**0.439**	**0.323**	**0.373**

To evaluate the effectiveness of components in TREAM, we performed ablation studies on the MIMIC-CXR dataset using LLaVA-Rad as the base architecture. The ablation results for each TREAM module are shown in Tables II and III, for NLG and clinical effectiveness metrics (CE), respectively, illustrating the contribution of each component to overall performance.

w/o Layer Feature Extraction(LFE): BLEU-4 decreases by 0.8% and F1 by 0.2%, indicating that single-layer features limit access to fine-grained visual information such as lesion boundaries and subtle density variations. This demonstrates the importance of capturing multi-level visual representations.

w/o Dual-weighted Temporal Fusion(DTF): BLEU-4 decreases by 0.7% and F1 by 0.1%. Equal-weight fusion of historical images fails to assign appropriate weights to more informative prior exams. The results highlight the value of adaptively emphasizing relevant longitudinal features based on semantic similarity and temporal proximity.

w/o Group Causal Transformer(GCT): BLEU-4 drops by 0.9% and F1 by 0.6%, the largest F1 decrease among temporal modules, highlighting the critical role of GCT in enforcing temporal causality and capturing longitudinal dependencies.

w/o Contrastive Learning Loss(CLL): BLEU-4 decreases by 1.0% and F1 by 0.3%, demonstrating that general text encoders lack radiology specific semantic understanding. These results indicate the necessity of maintaining accurate medical terminology.

w/o Chain-of-Medical-Thought(CoMT): BLEU-4 and F1 drop by 0.6% and 0.3%, respectively. End-to-end report generation without guided reasoning increases the risk of clinically inconsistent output and reduces report coherence.

F. Hyperparameter Analysis

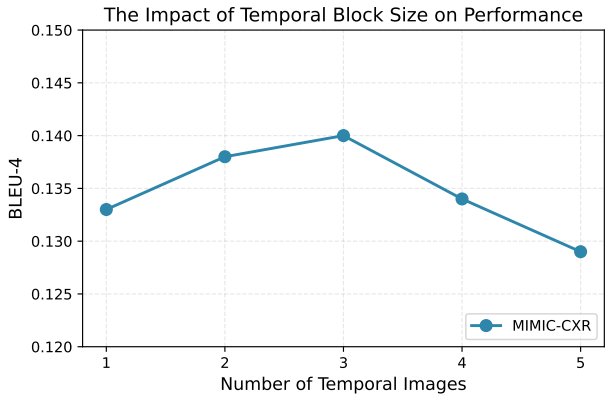


Fig. 2. Impact of temporal block size on BLEU-4 and F1 scores using 1–5 temporal images on MIMIC-CXR.

As shown in Figure 2, we further evaluate the effect of leveraging longitudinal information using different numbers of historical images. The results indicate that performance peaks at a temporal block size of 3, decreasing with fewer or more historical images.This trend reflects a balance between temporal context and information redundancy: too few images limit the ability of the model to capture disease progression, while too many may introduce outdated or irrelevant information. This demonstrates that the appropriate block size provides sufficient temporal context to generate accurate and clinically relevant radiology reports.

V. VISUAL COMPARISON OF GENERATED REPORTS

We compare TREAM generated reports with their corresponding ground truth reports, and the visualization results are shown in Figure 3. Different colors highlight distinct categories of medical findings: green indicates negative findings (e.g. “no focal consolidation,” “no effusion”), blue denotes normal anatomical descriptions (e.g. “normal,” “clear”), and red represents pathological terms (e.g. “effusion,” “pneumothorax,” “abnormal”).

As shown in the Figure 3, TREAM generates reports that are both clinically accurate and comprehensive. The model correctly captures key observations, including clear lungs, alongside well-demarcated cardiac and mediastinal contours, while also identifying additional clinically relevant details, such as bony structures without acute abnormalities and no free intraperitoneal gas detected. This improved descriptive completeness is supported by the Temporal Aggregator, which integrates multi-layer visual features with longitudinal context to capture disease evolution, while cross-modal alignment and step-wise diagnostic reasoning further ensure semantic accuracy and coherent, clinically reliable reports. Consequently, the generated reports more accurately reflect the current imaging findings and align closely with the ground truth report.

VI. CONCLUSION

In this paper, we propose TREAM, a temporal-aware MLLM that uses longitudinal imaging information to improve both the accuracy and consistency of radiology report generation. The model captures temporal dependencies across longitudinal images through Layer Feature Extraction to obtain multi-layer visual representations, combined with the Group Causal Transformer and Dual-weighted Temporal Fusion to fuse these features while preserving temporal causality and adaptively weighting prior exams based on relevant historical information, thereby tracking disease progression. In addition, it improves cross-modal alignment using a contrastive learning loss and supports step-wise diagnostic reasoning by employing Chain-of-Medical-Thought, allowing the model to maintain clinical accuracy and reduce hallucinations across longitudinal radiology tasks. Experiments on the MIMIC-CXR dataset show that our method achieves superior performance in natural language generation and clinical effectiveness metrics. Future work will explore extending the framework to other imaging modalities such as CT and MRI, incorporating complementary clinical information, and applying it to more diverse radiology tasks, thereby further improve performance.

REFERENCES

[1] Z. Chen, Y. Wang, J. Zhang *et al.*, “Collaboration between clinicians and vision–language models in radiology report generation,” *Nature Medicine*, vol. 30, pp. 1–10, 2024.

[2] T. Tu, S. Krishnan, N. Martinez *et al.*, “A multimodal multidomain multilingual medical foundation model for zero-shot clinical diagnosis,” *npj Digital Medicine*, vol. 8, no. 1, p. 32, 2025.

[3] J. V. Rawson, D. Smetherman, and E. Rubin, “Short-term strategies for augmenting the national radiologist workforce,” *American Journal of Roentgenology*, vol. 222, no. 4, p. e2430920, 2024.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[5] M. Li, W. Cai, K. Verspoor, S. Pan, X. Liang, and X. Chang, “Cross-modal clinical graph transformer for ophthalmic report generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 656–20 665.

[6] H. Wang, Y. Liu, X. Zhao *et al.*, “Kargen: Knowledge-enhanced automated radiology report generation using large language models,” in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2024, pp. 412–422.

[7] C. Li, J. Yang, Z. Chen *et al.*, “Med-gemini: Multimodal medical foundation models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024.

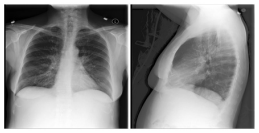
Input Radiology Image	Ground Truth Report	Our Model Generated Report
	There is no focal consolidation, pleural effusion, pneumothorax, or pulmonary edema identified. The heart size is normal. Mediastinal contours are normal.	No previous studies for comparison. There is no evidence of acute cardiopulmonary disease. The heart is normal in size. The mediastinal and hilar contours are normal. The lungs are clear. There is no pneumothorax or pleural effusion. The visualized osseous structures are unremarkable.
	Comparison is made to prior study from 5 days ago. There is a moderate right-sided pleural effusion, increased in size compared to the prior examination. The left lung is clear. There is no pneumothorax. The cardiac silhouette is normal in size. The mediastinal and hilar contours are stable. No acute osseous abnormality is detected.	COMPARISON: Chest radiograph from 5 days prior. FINDINGS: RIGHT PLEURAL EFFUSION: Moderate right-sided pleural effusion is present, demonstrating interval increase in volume compared to the prior study where it appeared small. The effusion obscures the right costophrenic angle. LEFT LUNG: The left lung remains clear with no focal consolidation. HEART AND MEDIASTINUM: The cardiac silhouette is normal in size. The mediastinal contours are stable. OTHER FINDINGS: No pneumothorax is identified. The visualized osseous structures demonstrate no acute fracture. IMPRESSION: 1. Moderate right-sided pleural effusion, increased from prior examination. 2. Clear left lung with no acute abnormality.

Fig. 3. Visual comparison of generated and ground truth reports. Colors indicate different categories of medical information.

[8] Y. Zhang, L. Sun, M. Chen *et al.*, "Radalign: Advancing radiology report generation with vision–language concept alignment," in *Proc. MICCAI*, 2025, to appear.

[9] Z. Li, Y. Chen, H. Wang *et al.*, "Instruction-guided fusion of multi-layer visual features in large vision-language models," *arXiv preprint arXiv:2501.08443*, 2024. [Online]. Available: https://arxiv.org/abs/2501.08443

[10] J. Liu, K. Huang, R. Wang *et al.*, "Ddgip: Radiology report generation through disease description graph and informed prompting," in *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[11] C. Pellegrini, E. Ozsoy, B. Busam, N. Navab, and M. Keicher, "Radialog: A large vision-language model for radiology report generation and conversational assistance," Nov. 2023, arXiv preprint arXiv:2311.18681. Available: https://arxiv.org/abs/2311.18681.

[12] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2gengpt: Radiology report generation with frozen llms," *Meta-Radiology*, vol. 1, no. 1, p. 100007, 2023.

[13] Z. Chen *et al.*, "Chexagent: Towards a foundation model for chest x-ray interpretation," Jan. 2024, arXiv preprint arXiv:2401.12208. Available: https://arxiv.org/abs/2401.12208.

[14] C. Wu *et al.*, "Towards generalist foundation model for radiology by leveraging web-scale 2d and 3d medical data," *Nature Communications*, vol. 16, p. 7866, 2025.

[15] J. M. Zambrano Chaves *et al.*, "A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings," *Nature Communications*, vol. 16, p. 3108, 2025.

[16] Y. Li, Z. Wang, Y. Li, H. Li, and L. Zhou, "Hergen: Hierarchical encoding for radiology report generation with explicit temporal structure," *Medical Image Analysis*, vol. 95, p. 103209, 2024.

[17] F. Liu, X. Zhou, Z. Wang, L. Li, L. Wang, and L. Zhou, "Libra: A prior-aware framework for radiology report generation with historical information," *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2257–2269, Jun. 2024.

[18] 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)*, Vancouver, BC, Canada, Jun. 2023, pp. 3334–3343.

[19] J. Park, S. Kim, B. Yoon, J. Hyun, and K. Choi, "M4cxr: Exploring multitask potentials of multimodal large language models for chest x-ray interpretation," *IEEE Transactions on Neural Networks and Learning Systems*, 2025, early Access.

[20] L. Xu *et al.*, "Damper: Distilling and aligning medical concepts with hypergraph for radiology report generation," *Information Fusion*, vol. 99, p. 101899, 2023.

[21] F. Pérez-García *et al.*, "RAD-DINO: Exploring scalable medical image encoders beyond text supervision," Jan. 2024, arXiv preprint arXiv:2401.10815. Available: https://arxiv.org/abs/2401.10815.

[22] J. M. Zambrano Chaves *et al.*, "Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation," 2024, arXiv preprint arXiv:2403.08002.

[23] A. E. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," Jan. 2019, arXiv preprint arXiv:1901.07042. Available: https://arxiv.org/abs/1901.07042.

[24] Z. Chen, A. Hernández Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut, "Meditron-70b: Scaling medical pretraining for large language models," *arXiv preprint arXiv:2311.16079*, 2023.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[26] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[28] A. Smit *et al.*, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," Apr. 2020, arXiv preprint arXiv:2004.09167.

[29] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," Oct. 2020, arXiv preprint arXiv:2010.16056.

[30] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," Apr. 2022, arXiv preprint arXiv:2204.13258.

[31] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 563–579.

[32] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10 578–10 587.

[33] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023.