sciendo

JAISCR, 2026, Vol. 16, No. 1, pp. 55 – 72
10.2478/jaiscr-2026-0003

CHESTXGEN: DYNAMIC MEMORY-AUGMENTED VISION-LANGUAGE TRANSFORMER WITH CONTEXT-AWARE GATING FOR RADIOLOGY REPORT GENERATION

Sharofiddin Allaberdiev[1], Asif Khan[2,3], Sardor Mamarasulov[4], Xiaojun Chen[1,*]

[1]*Shenzhen University*
China

[2]*University of Innovation Technologies*
Uzbekistan

[3]*Al-Sahli Medical Center*
Saudi Arabia

[4]*East China Normal University*
China

[*]*E-mail: xjchen@szu.edu.cn*

Submitted: 20th May 2025; Accepted: 22nd August 2025

Abstract

Chest X-ray analysis is vital for clinical screening, diagnosis, and treatment planning. The increasing workload on radiologists calls for robust automated solutions to generate accurate and standardized reports. Conventional report generation models often struggle to detect rare and anomalous diseases, particularly when faced with imbalanced datasets, which can compromise diagnostic knowledge accuracy. To address these limitations, we propose ChestXGen, a novel multimodal framework for automated radiology report generation. Our model is based on a fully Transformer-based encoder-decoder architecture that integrates Memory Augmented Transformer (MAT) blocks with a Context-Aware Bi-Gate (CABG) mechanism. These enable the model to capture long-range dependencies, effectively fuse visual and textual features, and better handle underrepresented conditions. Visual features are extracted using a ResNet-101-V2 backbone and refined through a shared memory module that continuously reinforces cross-modal associations. This integrated approach facilitates the generation of comprehensive, accurate, and contextually coherent reports. Extensive evaluation on the large-scale MIMIC-CXR dataset, comprising 377,110 images and corresponding free-text reports demonstrate that ChestXGen outperforms previous models on BLEU-1, BLEU-2, BLEU-3, and METEOR metrics. The results demonstrate the efficacy of Transformer-based models in substantially reducing radiologists' reporting burden while concurrently enhancing the precision and reliability of diagnostic interpretations.

Keywords: chest X-ray analysis, radiology report generation, memory augmented transformer, context-aware BiGate, anomalous diseases

1 Introduction

In clinical practice, analyzing chest radiographs and generating diagnostic reports are essential for accurate diagnosis, treatment planning, and patient management. Chest X-ray imaging is a cornerstone of clinical diagnostics, yet the manual generation of radiology reports is increasingly unsustainable due to rising imaging volumes and limited radiological expertise. Despite its importance, manual interpretation of chest X-rays is labor-intensive and prone to variability, potentially compromising diagnostic reliability and patient care. Automated systems that generate accurate and standardized reports are urgently needed to alleviate this burden and ensure consistent, high-quality outcomes. However, manual report generation is both cognitive load and labor-intensive for radiologists. Furthermore, handwritten reports can be difficult for patients and other healthcare providers to understand [1]. Therefore, automatically generating free-text radiological reports from chest X-rays is crucial, as it can alleviate the burden on radiologists, enhance the usability of computer-generated texts, and simplify the interpretation of results [2].

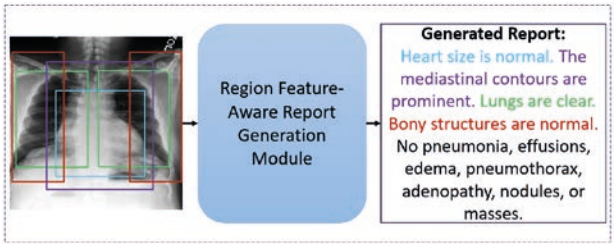


Figure 1. The proposed ChestXGen model detects key anatomical regions in chest X-rays where aligned visual and textual features are distinguished in different colors.

Several researchers have produced promising outcomes in the development of automated radiology report generation using various methods [3, 4]. Traditional approaches to automated report generation have predominantly employed convolutional neural networks (CNNs) for feature extraction combined with recurrent neural networks (RNNs) for text generation. Although these methods laid the groundwork for bridging image analysis with natural language processing, they struggle to capture long-term contextual dependencies and effectively align spatial image features with sequential textual information [5]. For example, these models may struggle to accurately detect and describe subtle abnormalities, such as early-stage pneumonia, leading to incomplete or imbalanced reports for certain diseases.

Recent advancements in transformer-based models have significantly improved the modeling of long-range dependencies and the integration of multimodal data. Leveraging self-attention mechanisms, transformers enable parallel processing and enhanced feature fusion, outperforming traditional RNN-based systems [6]. However, state-of-the-art transformer approaches fall short in addressing dynamic memory integration and precise cross-modal alignment, key challenges for generating comprehensive radiology reports. Building on the strengths of transformer architectures while overcoming these limitations, we propose ChestXGen, a novel framework that integrates Memory Augmented Transformer (MAT) blocks and a Context-Aware BiGate (CABG) mechanism. In ChestXGen, a ResNet-101-V2 backbone extracts robust visual features from chest X-rays, which are then processed by transformer layers with a shared memory matrix. The CABG mechanism refines region-specific features by filtering extraneous details and emphasizing diagnostically relevant cues, ensuring accurate alignment between visual and textual modalities and enabling the generation of structured, multi-sentence reports. We validate ChestXGen's performance through extensive ablation studies and comparisons with state-of-the-art methods, including the CheXReport baseline. Experiments on the MIMIC-CXR dataset show that ChestXGen achieves superior performance across multiple evaluation metrics, highlighting its potential to reduce radiologist workload and enhance diagnostic accuracy.

Our primary contributions are:

- *Introduction of ChestXGen:* A fully transformer-based model leveraging MAT blocks to capture long-range dependencies and improve cross-modal alignment, addressing the shortcomings of traditional architectures.
- *Context-Aware BiGate (CABG) Mechanism:* A novel component that dynamically adjusts feature representations, enhancing focus on diag-

nostically relevant image regions and improving report accuracy.

– *Comprehensive Validation:* Rigorous evaluation on the MIMIC-CXR dataset demonstrates ChestXGen's state-of-the-art performance, surpassing models like CheXReport [7] and confirming its ability to generate detailed, clinically relevant reports.

ChestXGen's enhanced cross-modal alignment and dynamic memory integration yield radiology reports with greater diagnostic precision on the MIMIC-CXR dataset. This improved performance can minimize diagnostic errors, optimize clinical workflows, and facilitate faster, more reliable clinical decision-making. The remainder of this paper is organized as follows: The second section reviews related work in automated image captioning and radiology report generation; the third section details the proposed ChestXGen architecture and methodology; the fourth section presents experimental evaluations and ablation studies; and the fifth section concludes the article with discussions on clinical implications and future research directions.

2 Relted work

In this section, we review previous efforts in image captioning and radiology report generation, highlighting the evolution from traditional feature extraction methods to advanced deep learning techniques for modeling complex intermodal relationships.

2.1 Image Captioning

Image captioning aims to generate descriptive text for images by capturing both objects and their interactions. Traditional image-based methods relied on manually engineered features (e.g. binary patterns and morphological descriptors) combined with classical machine learning algorithms [8]. However, these approaches struggled to generalize due to the inherent complexity of visual data [9]. Recent advances have primarily focused on encoder–decoder architectures [10]. In clinical contexts, radiological image analysis has inspired researchers to adopt similar frameworks, given the need to describe anatomical and pathological features accurately [11, 12, 13]. Most image captioning models employ CNNs to extract visual features, followed by recurrent neural networks (RNNs) or Transformer-based decoders for text generation [14, 15]. Alternative strategies have enhanced CNN-extracted representations by refining decoder architectures with Transformer blocks [16, 17]. Other studies have incorporated contrastive attention mechanisms to compare abnormal images with normal references, thus emphasizing critical regional features [18]. Moreover, some works adopt a sequential generation approach using Transformer decoders to produce multi-sentence narratives [19], while others utilize sparse attention to select the most salient visual features for improved semantic coherence [20].

2.2 X-ray Report Generation

The task of generating radiology reports from chest X-rays has seen significant progress through the application of traditional machine learning, deep learning, GANs, and attention mechanisms. Despite achieving promising results, many approaches still fall short in capturing the nuanced relationships between image regions and their corresponding textual descriptions [21]. Common limitations include difficulties in extracting fine-grained visual details, aligning image features with complex textual findings, and generating coherent multi-sentence narratives. Several studies have reported that existing methods, while effective in certain aspects, are limited by their inability to learn robust cross-modal associations or incorporate dynamic memory mechanisms [22, 23]. To address these challenges, recent work has explored fully Transformer-based architectures, which have proven effective in tasks such as language modeling and machine translation [24]. The architectural optimizations using simulated annealing [25] have demonstrated potential for improving transformer efficiency, but lack medical domain adaptation. In parallel, recent efforts have explored reinforcement learning-based approaches also for medical imaging tasks, such as novelty classification in cervical cancer detection [26] demonstrating the expanding use of adaptive learning in clinical diagnosis beyond report generation. In particular, the CheXReport model uses Transformers with ResNet-101-V2 and Swin Transformer blocks employing default

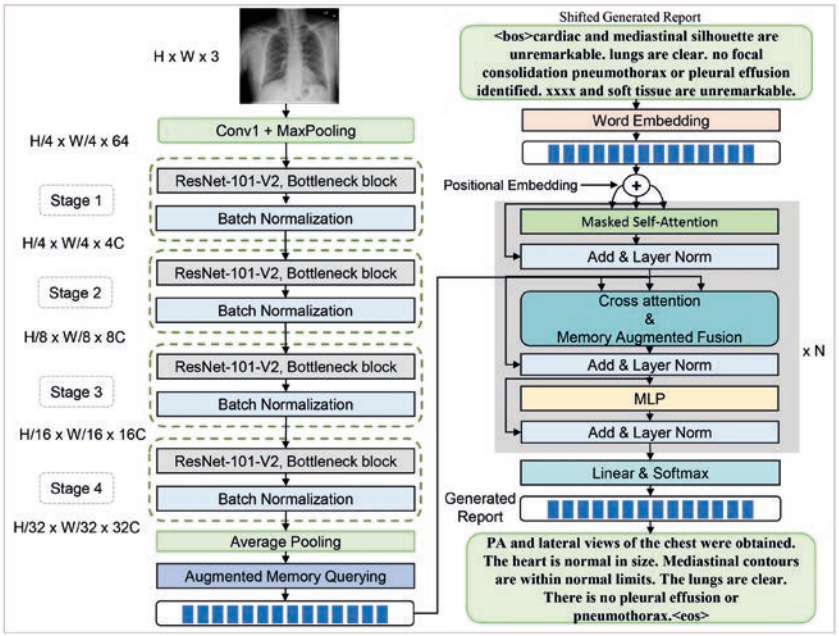


Figure 2. An overview of the proposed ChestXGen model. The X-ray image is processed through ResNet-101-V2 to extract visual features, followed by Augmented Memory Querying (AMQ) for feature aggregation. The Transformer decoder refines these features using self-attention and cross-attention with Memory Augmented Fusion as MAT via CABG to generate structured reports.

self-attention mechanisms [7]. However, these models still face issues such as hallucination and inadequate long-term context preservation. In contrast, our proposed model, ChestXGen, utilizes a MAT that incorporates a CABG mechanism. By integrating a shared memory module, our approach enables dynamic fusion of visual and textual features and ensures that the model can access its previous memories more accurately. This design improves cross-modal alignment and report generation quality, offering a promising direction to overcome the limitations observed in earlier works.

Recent benchmarks such as GLoRIA [27], BioViL-T [28], and MedPaLM [29] have advanced multimodal understanding in the medical domain. However, they do not achieve competitive performance on following generation metrics such as BLEU, METEOR, and ROUGE-L. GLoRIA focuses on region-text alignment using hierarchical contrastive learning, yet lacks a decoder for structured report generation. BioViL-T emphasizes retrieval and classification via contrastive learning but does not support memory-aware generative decoding. MedPaLM, while powerful for question answering and clinical reasoning, operates at a much larger scale and is not optimized for generating detailed multi-sentence radiology reports. In contrast, our ChestXGen model combines memory-augmented querying with per-token bidirectional-gated integration via CABG, enabling fine-grained semantic alignment and long-term context retention, and achieves strong results on BLEU, METEOR, and ROUGE-L metrics.

3 Material and Methods

In this section, we detail the key components of the proposed ChestXGen model. Unlike many conventional architectures that rely on standard encoder–decoder frameworks, our approach leverages a fully Transformer-based architecture augmented with MAT blocks via a CABG mechanism. This design enables ChestXGen to extract intrinsic visual features from chest X-ray images, capture their interrelationships, and store this information in a shared memory module, thereby facilitating the dynamic fusion of visual and textual representations. Our work builds upon the R2GenCMN framework

[12] by modifying the original Transformer architecture to include the CABG mechanism, employing ResNet 101 V2 as the backbone for feature extraction, and adopting a Transformer design as described in [7]. This hybrid approach supports deeper feature extraction and enhanced long-term context preservation via memory modules. We compare the performance of ChestXGen with state-of-the-art models to demonstrate its effectiveness in generating accurate and detailed radiology reports.

3.1 Dataset

We evaluate ChestXGen on the MIMIC-CXR benchmark dataset [30], which comprises 377,110 chest radiographs and 227,835 corresponding free-text reports. The reports are generally descriptive, with average lengths of 53.0 words in training, 53.1 words in validation, and 66.4 words in testing. Only reports that include the "Findings" section are used, following the official dataset partitioning as described in [31]. The large scale and complexity of MIMIC-CXR present significant challenges, yet also provide a diverse learning environment that enhances the model's ability to generate clinically relevant, well-structured reports.

3.2 ChestXGen Model

In this section, we explore in detail the task of generating radiology reports from image features, treating it as an image captioning task. Several viable solutions have been proposed in this similar area [32, 33]. Because images are usually inherently organized in a two-dimensional spatial format, we adhere to the standard sequence-to-sequence paradigm, as uniquely employed in [12] for this task. ChestXGen is formulated as a sequence-to-sequence problem where a chest X-ray image $\mathbf{I}$ is first transformed into a sequence of visual features, which are then decoded into a radiology report $\mathbf{Y}$. As follows $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s, ..., \mathbf{x}_S\}$, where $\mathbf{x}_s \in \mathbb{R}^d$, represent the sequence of visual features extracted from $\mathbf{I}$ via a ResNet 101 V2-based visual extractor by followed Augmented Memory Querying (AMQ), and let $\mathbf{Y} = \{y_1, y_2, ..., y_t, ..., y_T\}$, where $y_t \in \mathbb{V}$, denote the generated token sequence from the vocabulary $\mathbb{V}$, and T is the length of the report. The entire generation process is formalized as a recursive application of the chain rule, allowing the model to generate each token by conditioning on the previously generated tokens and the visual features

$$p(\mathbf{Y}|\mathbf{I}) = \prod_{t=1}^{T} p(y_t|y_1, ..., y_{t-1}, \mathbf{I}) \tag{1}$$

The model is then trained to maximize $p(\mathbf{Y}|\mathbf{I})$ through the negative conditional log-likelihood of $\mathbf{Y}$ given the $\mathbf{I}$, where the objective is to minimize the difference between the predicted and actual token sequences during training:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{T} \log p(y_t|y_1, ..., y_{t-1}, \mathbf{I}; \theta) \tag{2}$$

where θ is the parameters of the model.

An overview of the proposed ChestXGen model is illustrated in Figure 2, highlighting the role of memory augmented fusion through the CABG mechanism in generating structured radiology reports. This fusion process enhances the alignment between visual features from chest X-rays and their corresponding textual descriptions, ensuring more coherent and clinically relevant report generation. The following subsections provide a detailed explanation of the key components of the model and their contributions to improving the accuracy of the radiology report.

Our model demonstrates several distinct architectural innovations are illustrated in Table 1, including token-wise memory gating, context-aware bidirectional integration, and dynamic per-step adaptation, which are not present in prior works. These elements collectively improve the model's ability to maintain report coherence, especially in complex diagnostic narratives.

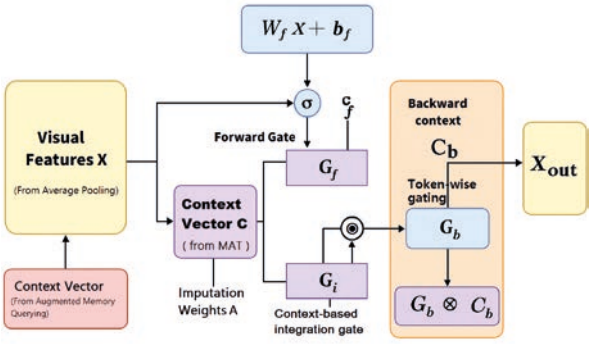


Figure 3. Schematic diagram of the proposed CABG module: visual features are refined through context vectors. Integration occurs via $\mathbf{C}$ (from MAT), forward $\mathbf{G}_f$, integration $\mathbf{G}_i$, and backward gates $\mathbf{G}_b$ before generating $\mathbf{X}_{\text{out}}$.

Table 1. Comparison of memory design and feature integration strategies between the proposed ChestXGen model and baseline architectures. Approaches include Cross-Modal Memory Networks (CMN), Dynamic Contrastive Learning (DCL) and Prompt-guided CheXReport, while ChestXGen employs the Context-Aware BiGate (CABG) mechanism.

Model Category	Architecture Type	Memory Usage, Gating	Model Adaptivity	Integration Level
CMN [12]	Shared global memory with decoder attention	No gating, fixed memory lookup during decoding	Low (static retrieval)	Global memory access from decoder
DCL [34]	Dual-encoder with graph-enhanced contrastive learning	No explicit memory, graph-based semantic alignment	Medium (relation-aware alignment)	Latent alignment across dual encoders
CheXReport [7]	Prompt-guided Transformer with dual-stage decoding	Implicit prompt-based memory, decoder fixed	Medium (retrieved prompts per case)	Retrieval-integrated decoding
ChestXGen (Ours)	Transformer decoder with token-wise local-global gating	Explicit per-token memory + bidirectional gating	High (dynamic at each decoding step)	Contextual local-global gating within decoder layers

3.3 Encoder

The encoder forms a critical component of the ChestXGen model, tasked with processing visual feature inputs and converting them into intermediate representations that serve as a foundation for generating textual reports [35, 36]. Built on a standard Transformer architecture, it leverages multi-head self-attention and positional encodings to capture complex visual patterns [37, 38]. Memory responses from the MAT module further enrich these representations. Let the memory-enhanced visual features be denoted as $\{\mathbf{r}_{\mathbf{x}_1}, \mathbf{r}_{\mathbf{x}_2}, ..., \mathbf{r}_{\mathbf{x}_S}\}$ where $\mathbf{r}_{\mathbf{x}_S}$ represents the memory-informed embedding of the s-th visual patch. The encoder transforms these inputs into a sequence of intermediate states $z_1, z_2, \ldots, z_S$ as follows:

$$\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_S\} = f_e(\mathbf{r}_{\mathbf{x}_1}, \mathbf{r}_{\mathbf{x}_2}, ..., \mathbf{r}_{\mathbf{x}_S}) \tag{3}$$

where $f_e(\cdot)$ is the encoder function comprising stacked layers of Transformer. Each layer applies multi-head self-attention and feed-forward operations to gradually refine the input representations, capturing both local details and global contextual information.

3.4 Visual Extractor

Producing image-based reports require extracting visual features from radiological images to ensure that critical diagnostic information is accurately captured, which then facilitates the generation of comprehensive and meaningful reports. For visual feature extraction, we employ ResNet-101-V2 pretrained on ImageNet. The input image is passed through four progressive stages of bottleneck blocks, where each stage includes batch normalization layers to stabilize training. The feature maps undergo convolutional and pooling operations, progressively reducing spatial dimensions while increasing channel depth. The final extracted visual feature representations have 512 dimensions per feature vectors. To enhance these extracted features, we integrate an AMQ step, which refines the alignment between visual feature maps and textual descriptions by retrieving relevant memory responses. In our method, the visual features of a radiology image $\mathbf{I}$, denoted as $\mathbf{X}$, are obtained through pre-trained CNNs such as ResNet [7]. Typically, an image is split into equal-sized patches (for example, 32 x 32 in ResNet), and their representations are obtained from the last convolutional layer. After extraction, these features are arranged into a sequence by concatenating them row by row across the image

patches. The resulting sequence then serves as the source input for all subsequent modules, formulated as:

$$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s, ..., \mathbf{x}_S\} = f_v(\mathbf{I}) \tag{4}$$

where $f_v(\cdot)$ refers to the visual extractor.

3.5 Decoder

The decoder complements the encoder by combining these intermediate visual representations $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_S\}$ with previously generated textual tokens. This allows it to produce meaningful and coherent radiology report sentences. The memory responses associated with textual tokens, denoted as $\{\mathbf{r}_{\mathbf{y}_1}, \mathbf{r}_{\mathbf{y}_2}, ..., \mathbf{r}_{\mathbf{y}_{t-1}}\}$ are fed into the decoder at each step. The output token y_t is then generated by conditioning on the encoder outputs and the memory-augmented textual features:

$$y_t = f_d(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_S, \mathbf{r}_{\mathbf{y}_1}, \mathbf{r}_{\mathbf{y}_2}, ..., \mathbf{r}_{\mathbf{y}_{t-1}}) \tag{5}$$

where $f_d(\cdot)$ represents the decoder function. This process repeats until the entire report is generated. This decoder integrates the CABG mechanism to fuse features dynamically and handles long-term dependencies via cross-attention, residual connections, and layer normalization.

3.6 Memory Augmented transformer (MAT) with Context-Aware BiGate (CABG)

To establish a connection between images and text, existing research typically aligns them by directly mapping their encoded representations. For instance, [39] used co-attention for this purpose. However, this approach frequently encounters the limitation that representations across modalities are difficult to align, requiring an intermediate medium to enhance and smooth the mapping process. To address this limitation, we propose using a MAT equipped with a CABG mechanism, which more effectively combines textual and visual features. This MAT integrates new input with past information, using BiGate to select relevant data from memory blocks and filter out extraneous details.

Memory Representation and Storage. Specifically, the MAT employs a matrix to preserve information for both the encoding and decoding processes, where each row of the matrix (i.e., a memory vector connecting images and texts). The MAT is designed to store, retrieve, and refine cross-modal information using a structured memory matrix:

$$M = \begin{bmatrix} m_1 \\ \vdots \\ m_N \end{bmatrix} \tag{6}$$

where $M \in \mathbb{R}^{N \times d}$ is the memory matrix containing N memory slots. Each memory vector $m_i \in \mathbb{R}^d$ stores multimodal associations between images and text. During the report generation process, the MAT operates in two main mechanisms: querying and responding, whose details are described as follows.

Memory Querying. We apply multi-threaded querying to perform this operation, where, in each thread number arbitrarily set in experiment. In querying the memory vectors, the first step is to ensure that the input visual and textual features are in the same representation space. Therefore, we convert each memory vector in $\mathbf{M}$ as well as the input features through a linear transformation as

$$\mathbf{k}_i = \mathbf{m}_i \cdot \mathbf{W}_k \tag{7}$$

$$\mathbf{q}_s = \mathbf{x}_s \cdot \mathbf{W}_q \tag{8}$$

$$\mathbf{q}_t = \mathbf{y}_t \cdot \mathbf{W}_q \tag{9}$$

where $\mathbf{W}_k$ and $\mathbf{W}_q$ are trainable weights for the conversion in the process. Next the model separately extract the most related memory vector to visual and textual corresponding features according to their distances D_{s_i} and D_{t_i} through, where the number of extracted memory vectors can be controlled by a hyperparameter $\mathcal{K}$ to regularize how much memory is used:

$$D_{s_i} = \frac{\mathbf{q}_s \cdot \mathbf{k}_i^\top}{\sqrt{d}} \tag{10}$$

$$D_{t_i} = \frac{\mathbf{q}_t \cdot \mathbf{k}_i^\top}{\sqrt{d}} \tag{11}$$

We refer to the retrieved memory vectors functioning as $\{\mathbf{k}_{s_1}, \mathbf{k}_{s_2}, ..., \mathbf{k}_{s_j}, ..., \mathbf{k}_{s_{\mathcal{K}}}\}$ and $\{\mathbf{k}_{t_1}, \mathbf{k}_{t_2}, ..., \mathbf{k}_{t_j}, ..., \mathbf{k}_{t_{\mathcal{K}}}\}$. Next, we calculate the importance of each memory vector in relation to the visual and textual features by normalizing all distances represented as

$$w_{s_i} = \frac{\exp(D_{s_i})}{\Sigma_{j=1}^{\mathcal{K}} \exp(D_{s_j})} \tag{12}$$

$$w_{t_i} = \frac{\exp(D_{t_i})}{\Sigma_{j=1}^{\mathcal{K}} \exp(D_{t_j})} \tag{13}$$

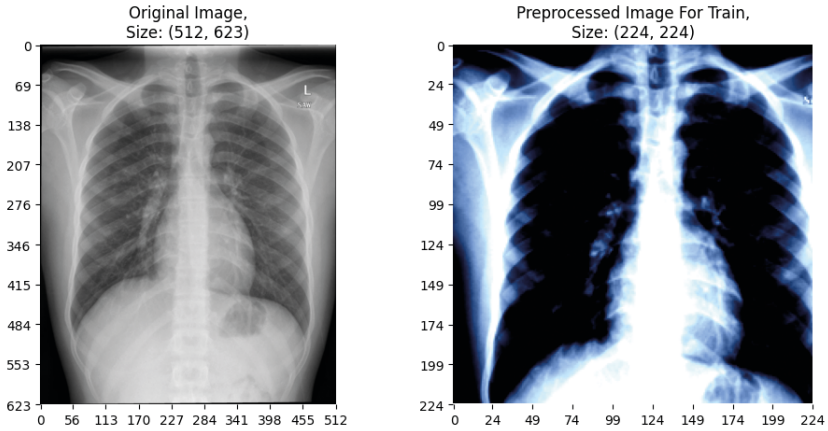


Figure 4. An example of the preprocessing applied to a chest X-ray image. The original image (left) is shown alongside the preprocessed version (right), which has been resized to 224×224 and enhanced through normalization and optional contrast adjustment for improved model training.

An important point is that these steps are applied in each thread to allow memory querying from different memory representation subspaces

Memory Responding. The response process follows a multi-threaded approach, aligning with the query process. In each thread, we begin by applying a linear transformation to the retrieved memory vector.

$$\mathbf{v}_i = \mathbf{m}_i \cdot \mathbf{W}_v \tag{14}$$

where $\mathbf{W}_v$ is the trainable weight associated with $\mathbf{m}_i$. Thus, all memory vectors $\{\mathbf{v}_{s_1}, \mathbf{v}_{s_2}, ..., \mathbf{v}_{s_j}, ..., \mathbf{v}_{s_\mathcal{K}}\}$ are conducted into $\{\mathbf{v}_{t_1}, \mathbf{v}_{t_2}, ..., \mathbf{v}_{t_j}, ..., \mathbf{v}_{t_\mathcal{K}}\}$. Next, we obtain the memory responses for the visual and textual features by weighting the transformed memory vectors with the weights obtained from the memory querying process:

$$\mathbf{r}_{\mathbf{x}_s} = \Sigma_{i=1}^{\mathcal{K}} w_{s_i} \mathbf{v}_{s_i} \tag{15}$$

$$\mathbf{r}_{\mathbf{y}_t} = \Sigma_{i=1}^{\mathcal{K}} w_{t_i} \mathbf{v}_{t_i} \tag{16}$$

where w_{s_i} and w_{t_i} represent the weights derived from the memory querying process. Just like in the querying step, the memory response process is executed across all threads to retrieve responses from various memory representation subspaces.

The CABG Mechanism. The CABG is a novel adaptive gating mechanism designed to dynamically refine feature representations by integrating both local and contextual information, as illustrated in Figure 3 and Algorithm **??**. Unlike conventional attention mechanisms, which assign only scalar attention weights to features, the BiGate adaptively modulates both forward and backward feature flows. This allows selective enhancement of informative features while suppressing redundant ones.

Given an input sequence of visual features $X \in \mathbb{R}^{T \times d}$, where T represents the sequence length and d denotes the feature dimension, the contextual attention mechanism assigns importance to the contextual memory representation $C \in \mathbb{R}^{T \times d}$, computed as:

$$A = \text{softmax}(W_c C) \tag{17}$$

where $W_c \in \mathbb{R}^{d \times 1}$ is a trainable parameter that projects the contextual features into a scalar attention weight, ensuring relevance-based weighting.

Using these attention scores, we derive the context-aware feature representations as:

$$\mathbf{C}_f = \mathbf{A}^T \mathbf{X} \tag{18}$$

$$\mathbf{C}_b = \mathbf{A}^T \mathbf{X}^{\text{rev}} \tag{19}$$

where $\mathbf{C}_f$ represents the forward contextual feature, and $\mathbf{C}_b$ denotes the backward contextual feature obtained by reversing the sequence. To regulate the incorporation of forward and backward contextual representations, CABG introduces two gated control functions:

$$\mathbf{G}_f = \sigma(\mathbf{W}_f \mathbf{X} + \mathbf{b}_f) \tag{20}$$

where $\sigma(\cdot)$ is the sigmoid activation function, ensuring that gating values remain within the range $[0,1]$. The weight matrices $\mathbf{W}_f, \mathbf{W}_b \in \mathbb{R}^{d\times d}$ are trainable parameters, while $\mathbf{b}_f, \mathbf{b}_b \in \mathbb{R}^d$ are the corresponding bias terms.

$$\mathbf{G}_b = \sigma(\mathbf{W}_b\mathbf{X} + \mathbf{b}_b) \tag{21}$$

Unlike prior co-attention, fixed gating methods, CABG dynamically computes both forward and backward gates conditioned on the contextual memory response and the decoding state. This bi-directional gating mechanism enables context-adaptive fusion by allowing each token to selectively attend to past information, thereby enhancing long-term dependency modeling. Such dynamic integration improves representation fidelity during report generation.

Algorithm 1. Context-Aware Bidirectional Gating (CABG): Token-Wise Feature Refinement via Adaptive Gating

Input: Visual feature vector $\mathbf{X}$, Context vector $\mathbf{C}$ from MAT
Output: Refined output vector $\mathbf{X}_{\text{out}}$

Forward Gate:
$\mathbf{c}_f \leftarrow \sigma(W_f\mathbf{X} + b_f)$
$\mathbf{G}_f \leftarrow \mathbf{X} \odot \mathbf{c}_f$
Context Integration Gate:
Compute imputation weights $\mathbf{A}$ from $\mathbf{X}, \mathbf{C}$
$\mathbf{G}_i \leftarrow \text{Fuse}(\mathbf{X}, \mathbf{C}, \mathbf{A})$
Backward Context Vector:
$\mathbf{C}_b \leftarrow \mathbf{G}_f + \mathbf{G}_i$
Backward Gating:
$\mathbf{G}_b \leftarrow \text{TokenGate}(\mathbf{C}_b)$
$\mathbf{X}_{\text{out}} \leftarrow \mathbf{G}_b \odot \mathbf{C}_b$
return $\mathbf{X}_{\text{out}}$

The final output representation is formulated as:

$$\mathbf{X}_{\text{out}} = \mathbf{G}_f \odot \mathbf{X} + \mathbf{G}_b \odot \mathbf{C}_b \tag{22}$$

where $\odot$ denotes element-wise multiplication. This enables the model to dynamically refine feature interactions while preserving local and global dependencies. This approach refines the transformer's self-attention by maintaining a balance between old and new features, reducing noise or redundancy. Unlike standard transformers, the MAT has an additional memory module with two main capabilities. Memory Slots allow the model to store and recall previously learned inferences, while a specialized Memory Read and Write Mechanism supports long-term associations within the transformer. This architecture enhances the mapping between visual and textual representations by introducing a structured memory module that facilitates long-term contextual retention.

3.7 Evaluation Metrics

The performance of our ChestXGen model is assessed using widely recognized natural language generation (NLG) metrics, including BLEU [40, 41]. These metrics are well-suited for assessing the semantic accuracy, fluency, and coherence of generated radiology reports compared to reference reports, ensuring that the model's output closely matches human-written text.

BLEU Score Calculation. BLEU (Bilingual Evaluation Understudy) is a widely used metric for evaluating automated text generation in various scenarios. It measures the quality of the generated text by comparing it with reference texts using n-grams. BLEU evaluates the precision of n-grams (contiguous sequences of n words) by comparing the generated text $\hat{\phi}$ to the reference text ϕ. Precision is calculated as the number of overlapping n-grams between the generated and reference texts divided by the total number of n-grams in the generated text:

$$P_n = \frac{\sum_{\text{n-grams}} \min(d_k(\hat{\phi}), d_k(\phi))}{\sum_{\text{n-grams}} d_k(\hat{\phi})} \tag{23}$$

where $\hat{\phi}$ represents the generated text, ϕ represents the reference text, and $d_k(\cdot)$ denotes the count of a specific n-gram.

Brevity Penalty (BP). BLEU applies a brevity penalty to prevent short generated texts from artificially inflating precision scores. The brevity penalty is defined as:

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp(1 - r/c), & \text{if } c \leq r \end{cases} \tag{24}$$

where c is the length of the generated text and r is the length of the reference text. If the generated text is shorter than the reference, the brevity penalty exponentially reduces the score.

Table 2. Ablation study for the contributions of the encoder and decoder in ChestXGen performance on the MIMIC-CXR dataset. Here, BL-n denotes the BLEU score using up to n-grams, MTR denotes METEOR, and RG-L denotes ROUGE-L.

DATA	MODEL	NLG METRICS					
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L
MIMIC -CXR	RESNET50-V2 + LSTM	0.238	0.128	0.019	0.013	0.117	0.245
	RESNET50-V2 + TRANSFORMER	0.258	0.139	0.094	0.028	0.135	0.249
	RESNET101-V2 + LSTM	0.241	0.129	0.087	0.026	0.124	0.247
	RESNET101-V2 + TRANSFORMER	0.263	0.147	0.103	0.041	0.138	0.257
	RESNET101-V1 + TRANSFORMER (MAT)	0.370	0.241	0.163	0.129	0.152	0.282
	RESNET101-V2 + TRANSFORMER (MAT)	**0.371**	**0.243**	**0.163**	**0.129**	**0.152**	**0.283**

Final BLEU Score. The final BLEU score combines the geometric mean of n-gram precision scores with the brevity penalty:

$$BLEU = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log P_n\right) \tag{25}$$

where N is the maximum n-gram order considered (e.g., $N = 4$ for BLEU-4), and P_n is the precision for each n-gram order.

METEOR Score Calculation. METEOR evaluates text quality using precision, recall, and additional features like stemming and synonym matching. It calculates a weighted harmonic mean of precision (P) and recall (R) at the unigram level:

$$METEOR = \frac{10 \cdot P \cdot R}{9P + R} \cdot (1 - p) \tag{26}$$

where P denotes precision, while R represents recall, both measured at the unigram level (single-word matches) between the predicted and ground-truth sentences. Additionally, a penalty factor (p) accounts for instances where the predicted and reference sentences share the same unigrams but differ significantly in structural organization:

$$p = 0.5 \cdot \left(\frac{c}{u_m}\right)^3 \tag{27}$$

where c is the number of matched chunks, and u_m represents the number of matched unigrams [42].

ROUGE-L Score Calculation. ROUGE-L focuses on the longest common subsequence (LCS) between the generated and reference reports. It calculates recall and precision based on the LCS and then computes the F-measure:

$$F_{\text{ROUGE-L}} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{28}$$

where β is typically set to 1 to give equal weight to precision and recall [43].

4 Results and discussion

To maximize the performance and generalizability of ChestXGen, a robust preprocessing pipeline was applied to the chest X-ray images from the MIMIC-CXR dataset. This pipeline standardizes input data by normalizing pixel intensities, resizing images while preserving anatomical proportions, and optionally enhancing contrast through adaptive histogram equalization, all of which are critical for reliable feature extraction and subsequent report generation. These preprocessing steps were designed to be fully compatible with the ResNet-101-V2 visual encoder used in our experiments.

Normalization and Standardization: Image normalization is performed to standardize pixel intensity distributions, ensuring uniformity across the dataset and facilitating effective feature learning by the pre-trained encoder. To accommodate the inherent variability in image dimensions, each chest X-

ray is resized to a fixed resolution of 224×224 pixels while preserving its aspect ratio. Zero padding is applied to the shorter axis to maintain the integrity of anatomical structures without distortion, as illustrated in Figure 4.

Contrast Enhancement: In addition to resizing and normalization, Contrast-Limited Adaptive Histogram Equalization (CLAHE) is optionally employed to enhance local contrast and improve the visibility of subtle radiological features. By mitigating noise and preserving fine details, CLAHE aids in capturing diagnostic information that is crucial for accurate report generation, particularly in images where subtle abnormalities are present.

4.1 Implementation Details

ChestXGen is implemented as a hybrid encoder–decoder system. The visual encoder (RESNET-101-V2) extracts high-dimensional feature representations from chest X-rays, which are subsequently processed through a contextual memory module (dimension: 2048) to achieve effective cross-modal alignment. The decoder is a Transformer-based network that comprises 4 layers with 16 attention heads and a hidden state dimension of 512. Beam search decoding (beam size = 5) is utilized during inference to optimize report fluency and coherence. The model is trained using the Adam optimizer as in [44] with an initial learning rate of 5×10^{-5}, decayed by 20% every 3 epochs $(\gamma = 0.85)$ over 50 epochs. Training was performed on an NVIDIA GeForce RTX 4090 GPU (CUDA Version 12.2), which facilitated efficient parallel processing and rapid convergence on the MIMIC-CXR dataset.

Model Size and Inference Efficiency. ChestXGen contains 59 million trainable parameters, with a total model size of 225.27 MB in float32 precision. To evaluate clinical feasibility, we measured inference time: on average, ChestXGen generates reports for a batch of seven chest X-ray images in 4.13 seconds, corresponding to approximately 0.59 seconds per image for end-to-end decoding. These results demonstrate that ChestXGen can operate in about real time on hospital-grade GPUs, supporting seamless integration into clinical workflows without introducing significant delays.

4.2 Optimization and Loss Function

We optimize ChestXGen using a token-level cross-entropy loss defined over the predicted report tokens. Formally, let $y = \{y_1, y_2, ..., y_T\}$ denote the ground truth tokens, and $\hat{y}_t$ the predicted distribution at timestep t. The objective function is:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p(\hat{y}_t = y_t) \tag{29}$$

This objective encourages the model to maximize the log-likelihood of generating the correct token at each decoding step. During training, we observe a steady reduction in training loss, from an initial value of 1.65 in the first epoch to 0.35 by epoch 50, indicating stable convergence.

Validation performance is measured using sequence-level language metrics (BLEU-1 to BLEU-4, METEOR, ROUGE-L). Since these metrics are non-differentiable and computed over complete generated reports, they are not directly comparable to the token-level training loss. Nonetheless, the steady improvement in these metrics over training epochs demonstrates that minimizing cross-entropy loss leads to better semantic and syntactic alignment with ground-truth reports.

4.3 Results (ablation study)

The ablation study was performed using established methodologies [7] to quantify the contributions of individual components within ChestXGen. We evaluated it based encoder–decoder configurations by varying the visual backbone and the decoding strategy. In particular, we compared models employing ResNet 50 V2 and ResNet 101 V2 backbones with LSTM-based and Transformer-based decoders. We compared our proposed MAT model with a CABG mechanism integrated into the Transformer architecture last. Our results show that Transformer-based decoders substantially outperform LSTM-based decoders in capturing long-range dependencies and generating coherent, detailed reports. Furthermore, among the variants using ResNet 101 V2 configuration as updated version of ResNet 101 V1 with MAT achieved the highest scores across all metrics. These findings indicate that the enhanced feature extraction provided by ResNet 101 V2, when combined with the dynamic memory augmentation and cross-modal attention

Table 3. Performance comparison with state-of-the-art methods on the MIMIC-CXR dataset. Here, BL-n refers to the BLEU-n score, MTR is METEOR, and RG-L is ROUGE-L. All scores for the state-of-the-art methods are directly cited from their original papers.

MODEL	BACKBONE	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
CA [45], (2021)	ResNet-50	0.350	0.219	0.152	0.109	0.151	0.283
CMCL [46], (2022)	ResNet-50	0.344	0.217	0.140	0.097	0.145	0.272
PGA [47], (2022)	ResNet101	0.356	0.222	0.151	0.111	0.140	0.280
CMN [12], (2022)	ResNet101	0.353	0.218	0.148	0.106	0.142	0.278
RM+MCLN [13], (2020)	ResNet101	0.353	0.218	0.145	0.103	0.142	0.277
DCL [34], (2023)	ViT	-	-	-	0.109	0.150	0.284
CheXReport [7], (2024)	Swin-B	0.354	0.225	0.145	0.127	0.147	**0.286**
ChestXGen (Ours)	ResNet-101 MAT	**0.371**	**0.243**	**0.163**	**0.129**	**0.152**	0.283

mechanisms of our Transformer encoder–decoder architecture, leads to more fluent and clinically accurate radiology reports.

The results indicate that the MAT architecture with the CABG mechanism improves both fluency and semantic accuracy, and that the ResNet 101 V2 backbone yields the best performance among all tested configurations. Consequently, the final model configuration adopts ResNet 101 V2 for superior visual feature extraction.

We now present an in-depth analysis of the overall performance of ChestXGen and examine the effects of our transformer-based encoder-decoder architecture augmented with MAT blocks and the CABG mechanism. Table 2 summarizes the quantitative results obtained from different configurations in the MIMIC-CXR data set. In our hybrid architecture, visual features are extracted using ResNet 101 V2 and processed by Transformer layers that incorporate both MAT and CABG modules. These modules ensure that critical visual information is retained and dynamically aligned with the generated text. The Transformer decoder processes input sequences of arbitrary length, which enables the generation of longer, more detailed reports as indicated by the higher BLEU scores. Moreover, the memory augmentation module preserves salient visual features and facilitates their effective integration during text generation. This is reflected in improved METEOR and ROUGE L scores, confirming that ChestXGen not only generates fluent and semantically rich reports but also meets clinical expectations. Our experiments demonstrate that integrating MAT with the CABG mechanism produces significant performance gains compared to baseline architectures using LSTM-based decoders or conventional memory networks.

4.4 Transformer Encoder and Decoder Effect

Our results reveal that Transformer-based decoders yield consistently higher evaluation scores compared to LSTM-based decoders. The Transformer decoder's ability to capture long-range dependencies and process input sequences of arbitrary length results in coherent and contextually relevant reports. In our architecture, the encoder extracts enhanced visual features via the ResNet 101 V2 backbone and passes these representations through Transformer layers that leverage self-attention, further reinforced by memory augmentation. This design leads to improvements in both BLEU and ROUGE L metrics, underscoring the effectiveness of our Transformer encoder–decoder framework in capturing complex radiological patterns and generating detailed narratives.

4.5 Modal Memory Effect

The dynamic memory mechanism integrated into ChestXGen is critical for aligning visual and textual modalities. Unlike traditional static memory networks, our MAT module is tightly integrated into the Transformer decoder. This integration al-

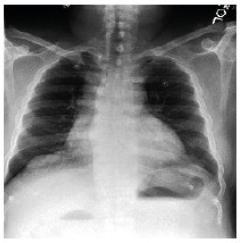
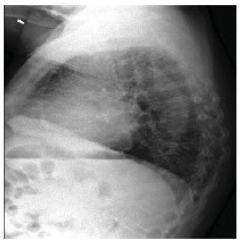
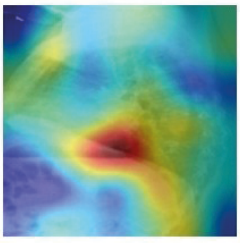
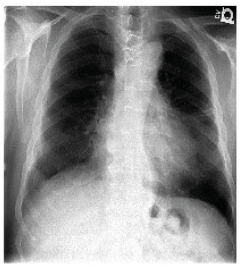
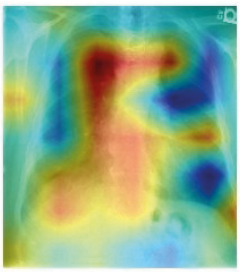
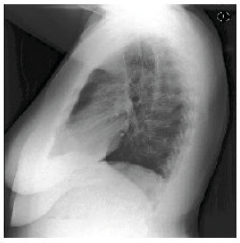
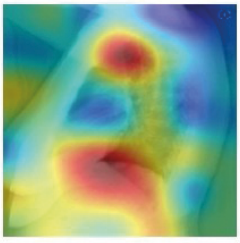
Image	Ground Truth	ResNet101-V1 + LSTM	ResNet101-V2 + Transformer	ResNet101-V2 + Transformer (MAT)	Aggregated Attention Map
	The heart size and mediastinal contours are prominent but similar to prior studies. the lungs are clear. there is no pleural effusion or pneumothorax.	No indication of new focal consolidation, and the lung fields remain clear.	Cardiac size remains normal, lungs are clear. Bony structures are normal. no pneumonia effusions edema pneumothorax adenopathy nodules or masses.	Heart size is normal. there has been no significant change. Mediastinal contours are stable with mild prominence. Lungs are clear without focal consolidation. No pleural effusion, pneumothorax, adenopathy, nodules, or masses. Bony thoracic structures are intact.	
	Prominent interstitial markings are again seen not significantly changed. there is no overt pulmonary edema. there is no pleural effusion. cardiomediastinal silhouette is stable. coronary artery calcifications and or stents are noted. chronic unk deformity in the lower thoracic spine.	Prominent interstitial markings remain unchanged. No signs of pneumothorax or pleural effusion. Pulmonary edema is not observed. Catheter in normal position.	Borderline heart size not changed. tortuous calcified aorta. no active pulmonary disease. The cardiomediastinal silhouette, artery are normal. Chronic deformity is lower thoracic spine. mild spondylosis.	Moderate bilateral interstitial markings are seen, not changed. No effusion is present. The cardiomediastinal silhouette remains stable. Coronary artery stents are noted. Chronic deformity is seen in the lower thoracic spine. No pneumothorax is identified.	
	Since the prior radiograph there has been no significant change. there is no focal consolidation pleural effusion pneumothorax or pulmonary edema. cardiomediastinal silhouette is unchanged and notable for tortuous aorta and mild cardiomegaly . median sternotomy wires are present and intact. clips are seen in the midline of the thorax . bony structures are intact.	Heart size is unchanged. No obvious signs of lung problems like fluid or collapse. Mediastinum and bony structures are seem fine. Sternal wires and clips are visible.	No focal airspace disease, effusion, or pneumothorax and Cardiomediastinal silhouette size is normal. mediastinal and hilar contours are unchanged. pulmonary vasculature is not engorged. lungs are clear and no pleural effusion is identified. no acute osseous abnormality is visualized.	Patient is status post median sternotomy and cabg with no significant change since prior radiograph. Heart size is within normal limits. The cardiomediastinal silhouette is stable, showing mild cardiomegaly and a tortuous aorta. No focal consolidation, pleural effusion, pneumothorax, or pulmonary edema is seen. Sternotomy wires and midline surgical clips are intact. Bony thoracic structures are unremarkable.	
	The cardiac, mediastinal and hilar contours appear stable. The heart is normal in size. There is no pleural effusion or pneumothorax. The lungs appear clear. The patient is status post anterior cervical fusion. Surgical clips project over the left upper quadrant. There has been no significant change.	The lungs are clear with no significant changes. The heart size appears slightly abnormal. The cardiomediastinal silhouette is stable. No pleural effusion or pneumothorax is identified.	Frontal and lateral radiographs of the chest show a stable cardiomediastinal silhouette. The heart is normal in size. Lungs are clear. Atherosclerotic calcification of the aortic arch is noted.	Structure of cardiac appears fine. The heart is normal. No pleural effusion or pneumothorax is present, no fluid or collapse noticed. Prior surgical involving cervical fusion is noted, with metallic artifacts in the left upper quadrant There is no significant change.	

Figure 5. Comparison of ground truth reports generated by the RESNET101-V1 + LSTM, RESNET101-V2 + TRANSFORMER, and RESNET101-V2 + TRANSFORMER (MAT) models for chest X-ray images selected from the MIMIC-CXR test set.

lows for continuous, dynamic interactions between the encoded visual features and the generated text, enabling the model to retain and update relevant information during decoding. The resulting improvements in BLEU scores and METEOR demonstrate that our approach enhances semantic accuracy and fluency by preserving key diagnostic details that are essential for generating clinically accurate reports.

4.6 Comparison with State-of-the-Art Models

We compared ChestXGen with several state-of-the-art models for radiology report generation using the MIMIC-CXR dataset. The compared approaches include Contrastive Attention (CA) [45], Competence-based Multimodal Curriculum Learning Framework (CMCL) [46], Prior Guided Attention (PGA) [47], Cross-Modal Memory Networks (CMN) [12], Relational Memory and Memory-Driven Conditional Layer Normalization (RM+MCLN) [13], Knowledge Graph with Dynamic Structure and Contrastive Learning (DCL) [34], and CheXReport [7]. Table 3 summarizes the performance of the models alongside ChestXGen.

Our ChestXGen model, particularly the variant utilizing ResNet 101 V2 as the visual encoder with a Transformer-based decoder, outperforms all competing methods across key evaluation metrics. The consistently high BLEU scores, alongside a slightly lower ROUGE-L, suggest that our model is effective in producing text with precise word choices and strong local coherence. The notably higher METEOR score further highlights our model's ability to generate semantically meaningful and fluent reports that closely align with the reference texts. This is particularly important in radiology, where clinical accuracy often depends more on semantic equivalence than exact word or phrase matching. While ROUGE-L can be helpful in capturing structural overlap, it may underrepresent semantically correct reports that use alternative phrasing. This limitation becomes more apparent in medical text generation, where variation in report templates and terminology is common.

The improvements are primarily attributed to two factors. First, the ResNet 101 V2 backbone extracts richer and more detailed visual features than its counterparts, which directly enhances the alignment with the textual descriptions. Second, the integration of the MAT with the CABG mechanism enables dynamic retention and alignment of critical visual information during decoding. This integrated memory mechanism improves the overall coherence and clinical relevance of the generated reports, effectively addressing the limitations of traditional static memory networks.

ChestXGen (ResNet 101 V2 +Transformer as MAT) demonstrates a significant advancement by achieving competitive performance with state-of-the-art models while delivering more detailed and semantically accurate descriptions. This validates our approach and underscores the benefits of combining advanced visual feature extraction with dynamic memory augmentation in a Transformer-based framework.

4.7 Qualitative Analysis

Qualitative analysis was conducted by comparing the generated reports and aggregated attention maps against ground truth annotations from the MIMIC-CXR dataset. In Table (Figure) 5, we present representative examples comparing reports generated by baseline models such as ResNet 101 V1 + LSTM, ResNet 101 V2 + Transformer and ResNet101-V2 + Transformer (MAT) with those produced by ChestXGen. Our observations indicate that LSTM-based decoders often generate shorter reports that miss critical details and struggle to capture subtle diagnostic features, such as coronary calcifications or surgical clips. In contrast, Transformer-based decoders perform much better at retaining essential information. Specifically, ChestXGen, which integrates a Transformer-based decoder with MAT and CABG, produces more detailed and accurate reports. The generated text not only demonstrates greater fluency and semantic coherence but also closely aligns with clinical narratives, enhancing the reliability of automated radiology reporting. Aggregated attention maps from the Transformer encoder's final layer further corroborates these findings. The attention maps of ChestXGen consistently concentrate on clinically significant regions, including the lungs, heart, and mediastinum. For instance, in one example, the attention field exhibits a broader focus corresponding to prominent mediastinal contours, which aligns with clinical observations. These qualitative results confirm that ChestXGen effectively integrates vi-

sual features with generated text, reduces hallucinations, and produces comprehensive radiology reports that adhere to medical standards.

5 Conclusion, Limitations and Future work

This paper introduced ChestXGen, a novel transformer-based framework for radiology report generation that incorporates Memory-Augmented Transformer (MAT) blocks and a Context-Aware BiGate (CABG) mechanism. Our approach dynamically fuses visual features from ResNet-101-V2 with textual representations via token-wise gating and memory-augmented cross-modal alignment. Extensive experiments on the large MIMIC-CXR dataset demonstrate strong performance, achieving BLEU-1 (0.371), BLEU-2 (0.243), BLEU-3 (0.163), BLEU-4 (0.129), and METEOR (0.152), thereby validating the model's capability to generate clinically meaningful and coherent reports while reducing diagnostic omissions and hallucinations. The CABG mechanism proves especially effective in maintaining long-range dependencies and focusing on diagnostically relevant visual features.

Despite these promising results, following limitations should be acknowledged. First, although CABG improves the fusion of memory and visual context knowledge, the use of memory modules increases the computational burden, which may pose challenges for deployment in low-power embedded systems. Second, while the model has been trained and evaluated on the large-scale MIMIC-CXR benchmark, its generalizability to other clinical domains or imaging modalities, especially those involving unseen diseases, remains uncertain and warrants further investigation.

Future work will focus on several key directions. Architecturally, we aim to enhance the CABG mechanism through multi-head gating strategies and fine-grained analysis of gate activations to improve both performance and interpretability. We also plan to incorporate model compression and quantization techniques to facilitate efficient deployment in resource-constrained environments. Beyond architectural improvements, future research will evaluate the generalizability of ChestXGen across external and updated datasets, extend its applicability to other imaging modalities and disease domains through transfer learning, and rigorously assess clinical quality in collaboration with expert radiologists. Additionally, we will develop advanced visualization tools and post-hoc interpretability analyses, investigate optimal human–AI collaboration strategies for clinical workflow integration, and systematically address ethical considerations related to bias, fairness, and equitable clinical deployment.

In conclusion, the combination of a ResNet 101 V2 backbone, Transformer-based encoder–decoder framework, and dynamic memory augmentation via MAT with the CABG mechanism enables ChestXGen to generate high-quality, clinically relevant radiology reports. These results open up new avenues for further research to advance the state-of-the-art in automated radiology report generation.

Acknowledgment

This research was supported in part by NSFC under Grant no. 92270122, 62476174 and 62201179; and in part by Natural Science Foundation of Guangdong Provincial under Grant no. 2023A1515012584; and in part by the Shenzhen Basic Research Project (Natural Science Foundation) Basic Research Key Project (NO. JCYJ20241202124430041).

References

[1] P. Roy, A. Bhunia, A. Das, P. Dhar, U. Pal, Keyword spotting in doctor's handwriting on medical prescriptions, Expert Systems with Applications 76 (2017) 113–128. https://doi.org/10.1016/j.eswa.2017.01.042, doi:10.1016/j.eswa.2017.01.042.

[2] Y. Han, G. Holste, Y. Ding, Radiomics-guided global-local transformer for weakly supervised pathology localization in chest x-rays, IEEE Transactions on Medical Imaging 42 (3) (2022) 750–761.https://doi.org/10.1109/TMI.2022.3217292, doi:10.1109/TMI.2022.3217292.

[3] C. Bluethgen, P. Chambon, J.-B. Delbrouck, R. van der Sluijs, M. Połacin, J. M. Zambrano Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, A. S. Chaudhari, A vision–language foundation model for the generation of realistic chest x-ray images, Nature Biomedical Engineering (2024) 1–13 https://doi.org/10.1038/s41551-024-01152-3, doi:10.1038/s41551-024-01152-3

[4] G. Reale-Nosei, E. Amador-Domínguez, E. Serrano, From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation, Medical Image Analysis (2024) 103264.

[5] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From show to tell: A survey on deep learning-based image captioning, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (1) (2022) 539–559. https://doi.org/10.1109/TPAMI.2022.3148210, doi:10.1109/TPAMI.2022.3148210

[6] Z. Tian, A. Liu, G. Zhu, X. Chen, A paralleled cnn and transformer network for ppg-based cuff-less blood pressure estimation, Biomedical Signal Processing and Control 99 (2025) 106741. https://doi.org/10.1016/j.bspc.2023.106741, doi:10.1016/j.bspc.2023.106741

[7] F. Zeiser, C. Costa, G. Gabriel R, A. Maier, R. da Rosa Righi, Chexreport: A transformer-based architecture to generate chest x-ray reports suggestions, Expert Systems with Applications 255 (2024) 124644. https://doi.org/10.1016/j.eswa.2023.124644, doi:10.1016/j.eswa.2023.124644

[8] N. Linna, C. E. Kahn Jr, Applications of natural language processing in radiology: A systematic review, International Journal of Medical Informatics 163 (2022) 104779.

[9] H. Sharma, D. Padha, A comprehensive survey on image captioning: From handcrafted to deep learning-based techniques, a taxonomy and open research issues, Artificial Intelligence Review 56 (11) (2023) 13619–13661. https://doi.org/10.1007/s10462-023-10489-5, doi:10.1007/s10462-023-10489-5

[10] Z. Wang, L. Wang, X. Li, L. Zhou, Diagnostic captioning by cooperative task interactions and sample-graph consistency, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).

[11] D. Singh, M. Kaur, J. M. Alanazi, A. A. AlZubi, H. N. Lee, Efficient evolving deep ensemble medical image captioning network, IEEE Journal of Biomedical and Health Informatics 27 (2) (2022) 1016–1025. https://doi.org/10.1109/JBHI.2022.3149312, doi:10.1109/JBHI.2022.3149312

[12] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal memory networks for radiology report generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5904–5914.

[13] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1439–1449.

[14] Y. Wang, J. Xu, Y. Sun, End-to-end transformer based model for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 12, 2022, pp. 2585–2594.

[15] W. M. da Silva, S. C. Cazella, R. S. Rech, Deep learning algorithms to assist in imaging diagnosis in individuals with disc herniation or spondylolisthesis: A scoping review, International Journal of Medical Informatics (2025) 105933.

[16] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 10578–10587.

[17] Y. Tang, H. Yang, L. Zhang, Y. Yuan, Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation, Expert Systems with Applications 237 (2024) 121442. https://doi.org/10.1016/j.eswa.2023.121442, doi:10.1016/j.eswa.2023.121442

[18] M. Lin, T. Li, Z. Sun, G. Holste, Y. Ding, F. Wang, Y. Peng, Improving fairness of automated chest radiograph diagnosis by contrastive learning, Radiology: Artificial Intelligence 6 (5) (2024) e230342. https://doi.org/10.1148/ryai.230342, doi:10.1148/ryai.230342

[19] V. D. Rao, B. N. Shashank, S. Nagesh Bhattu, Improved image captioning using gan and vit, in: International Conference on Computer Vision and Image Processing, Springer Nature Switzerland, 2023, pp. 375–385. https://doi.org/10.1007/978-3-031-38366-5_30, doi: 10.1007/978-3-031-38366-5_30.

[20] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, Y. Zou, Unify, align and refine: Multi-level semantic alignment for radiology report generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2863–2874.

[21] J. H. Moon, H. Lee, W. Shin, Y. H. Kim, E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, IEEE Journal of Biomedical and Health Informatics 26 (12) (2022) 6070–6080. https://doi.org/10.1109/JBHI.2022.3208800, doi:10.1109/JBHI.2022.3208800

[22] J. Duan, J. Xiong, Y. Li, W. Ding, Deep learning based multimodal biomedical data fusion: An overview and comparative review, Information Fusion (2024) 102536.

[23] S. Li, P. Qiao, L. Wang, M. Ning, L. Yuan, Y. Zheng, J. Chen, An organ-aware diagnosis framework for radiology report generation, IEEE Transactions on Medical Imaging (2024). https://doi.org/10.1109/TMI.2024.3361536, doi: 10.1109/TMI.2024.3361536

[24] H. Li, H. Wang, X. Sun, H. He, J. Feng, Context-enhanced framework for medical image report generation using multimodal contexts, Knowledge-Based Systems (2025) 112913.

[25] M. Trzciński, S. Łukasik, A. H. Gandomi, Optimizing the structures of transformer neural networks using parallel simulated annealing, Journal of Artificial Intelligence and Soft Computing Research 14 (3) (2024) 267–282, Społeczna Akademia Nauk. https://doi.org/10.2478/jaiscr-2024-0015, doi: 10.2478/jaiscr-2024-0015

[26] S. Muksimova, S. Umirzakova, K. Shoraimov, J. Baltayev, Y. I. Cho, Novelty classification model use in reinforcement learning for cervical cancer, Cancers 16 (22) (2024) 3782. https://doi.org/10.3390/cancers16223782, doi: 10.3390/cancers16223782

[27] X. Huang, Y. Zhang, J. P. Cohen, L. Zhang, E. P. Xing, Gloria: A multimodal global-local representation learning framework for medical vision-language tasks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3942–3951.

[28] B. Boecking, N. Usuyama, N. Bannur, C. Yu, S. Pournejat, A. Sethi, Z. Zhan, K. Lakhotia, A. Kumar, P. He, et al., Making the most of text semantics to improve biomedical vision–language processing, arXiv preprint arXiv: 2204.09817 (2022). http://arxiv.org/abs/2204.09817, arXiv: 2204.09817

[29] K. Singhal, S. Azizi, T.-J. Tu, S. Mahdavi, Y. Berne, J. Wei, H. W. Chung, N. Scales, et al., Large language models encode clinical knowledge, Nature 620 (7972) (2023) 172–180. https://doi.org/10.1038/s41586-023-06291-2, doi:10.1038/s41586-023-06291-2

[30] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, Y. Peng, S. Horng, Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, arXiv preprint arXiv: 1901.07042 (2019). http://arxiv.org/abs/1901.07042, arXiv: 1901.07042

[31] B. Jing, Z. Wang, E. Xing, Show, describe and conclude: On exploiting the structure information of chest x-ray reports, arXiv preprint arXiv: 2004.12274 (2020). http://arxiv.org/abs/2004.12274, arXiv: 2004.12274

[32] W. Ansar, S. Goswami, A. Chakrabarti, A survey on transformers in nlp with focus on efficiency, arXiv preprint arXiv: 2406.16893 (2024). http://arxiv.org/abs/2406.16893, arXiv: 2406.16893

[33] J. Wang, W. Jiang, L. Ma, W. Liu, Y. Xu, Bidirectional attentive fusion with context gating for dense video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7190–7198.

[34] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, X. Chang, Dynamic graph enhanced contrastive learning for chest x-ray report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3334–3343.

[35] D.-P. Kuo, Y.-C. Chen, S.-J. Cheng, K. L.-C. Hsieh, Y.-T. Li, P.-C. Kuo, Y.-C. Chang, C.-Y. Chen, A vision transformer-convolutional neural network framework for decision-transparent dual-energy x-ray absorptiometry recommendations using chest low-dose ct, International Journal of Medical Informatics 199 (2025) 105901.

[36] Z. Hu, A. Iscen, C. Sun, Z. Wang, K. W. Chang, Y. Sun, C. Schmid, D. A. Ross, A. Fathi, Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23369–23379.

[37] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys (CSUR) 51 (6) (2019) 1–36. https://doi.org/10.1145/3241036, doi: 10.1145/3241036

[38] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077–6086.

[39] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2023, pp. 2787–2797.

[40] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[41] G. Datta, N. Joshi, K. Gupta, Analysis of automatic evaluation metric on low-resourced language: Bertscore vs bleu score, in: International Conference on Speech and Computer, 2022, pp. 155–162.

[42] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[43] C. Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[44] D. P. Kingma, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014). http://arxiv.org/abs/1412.6980, arXiv: 1412.6980

[45] F. Liu, S. Ge, Y. Zou, X. Wu, Competence-based multimodal curriculum learning for medical report generation, arXiv preprint arXiv:2206.14579 (2022). http://arxiv.org/abs/2206.14579, arXiv: 2206.14579

[46] F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, X. Sun, Contrastive attention for automatic chest x-ray report generation, arXiv preprint arXiv: 2106.06965 (2021). http://arxiv.org/abs/2106.06965, arXiv: 2106.06965

[47] B. Yan, M. Pei, M. Zhao, C. Shan, Z. Tian, Prior guided transformer for accurate radiology reports generation, IEEE Journal of Biomedical and Health Informatics 26 (11) (2022) 5631–5640. https://doi.org/10.1109/JBHI.2022.3181343, doi: 10.1109/JBHI.2022.3181343.

Xiaojun Chen is a professor at the College of Computer Science and Software Engineering, Shenzhen University. He received his Ph.D. from the Harbin Institute of Technology. He is a distinguished member of CCF, and Deputy Chair of the CCF YOCSEF Headquarters. His research interests include deep learning, large-scale knowledge modeling, and multimodal big data analysis.
https://orcid.org/0000-0002-2818-4652

Sharofiddin Allaberdiev is currently a Ph.D. candidate at Computer Science and Software Engineering, Shenzhen University under the supervision of Prof. Xiaojun Chen. He received his Master's degree in Computer Science from WTU, China. He was a visiting scholar at the Indian Institute of Science and Technology in 2022. He served as a Lecturer at Korean Ajou University in Tashkent, until 2023. His research interests include radiology report generation, and multimodal data fusion.
https://orcid.org/0000-0001-6087-8024

Asif Khan is a researcher specializing in data science, NLP, social network analysis, and multimodal learning. His work focuses on analyzing human behavior through computational methods, including social media analysis, misinformation detection, and computational psychology. He holds a Ph.D. in Computer Science and Technology and a Master's in Software Engineering from Beijing Institute of Technology, as well as a Master's in Computer Science from the University of Peshawar.
https://orcid.org/0000-0002-5863-8467

Sardor Mamarasulov has graduated a Ph.D. from the School of Computer Science and Technology, East China Normal University, Shanghai, China. He received his bachelor's degree in Information Technologies from Tashkent University of Information Technologies in 2014 and his master's degree in Software Engineering from the Beijing Institute of Technology in 2018. His research focuses on machine learning, pattern recognition, and computer vision, with an emphasis on deepfake detection technologies.
https://orcid.org/0009-0003-6481-3932