
Proceedings Paper:
Khandeparker, A. and Lu, P. orcid.org/0000-0002-0199-3783 (Accepted: 2025) Transform(AI)ng Radiology with CheXSBT: Integrating Dual-Attention Swin Transformer with BERT for Seamless Chest X-Ray Report Generation. In: Ali, S. and Hogg, D.C., (eds.) Medical Image Understanding and Analysis. Medical Image Understanding and Analysis (MIUA) 2025, 15-17 Jul 2025, Leeds, UK. Lecture Notes in Computer Science (LNCS) . Springer Nature ISBN 978-3-031-98687-1 (In Press)

Transform(AI)ng Radiology with CheXSBT: Integrating Dual-Attention Swin Transformer with BERT for Seamless Chest X-Ray Report Generation

Aradhya Khandeparker[0009−0002−6920−5517] and Ping Lu[0000−0002−0199−3783]

School of Computer Science, University of Leeds, Leeds, LS2 9JT, United Kingdom
lmfx5605@leeds.ac.uk, P.Lu@leeds.ac.uk

Abstract. Radiology reports are crucial for diagnosing diseases, yet generation them is time-consuming, places a significant workload on medical professionals, and is subject to inter-expert variability, as different radiologists may interpret the same X-ray differently. This paper presents a novel hybrid AI model called CheXSBT, which combines our custom-designed Dual-Attention Swin Transformer (DAST) for vision processing with BERT for natural language understanding to automate the generation of chest X-ray (CXR) reports. Leveraging the MIMIC-CXR dataset, which includes over 370,000 X-ray images and their corresponding reports, CheXSBT learns to interpret chest X-ray images and convert them into structured, meaningful text. Our study focuses on two main objectives: (1) automating report generation to accelerate the diagnostic process and (2) improving model interpretability to foster trust among radiologists. The approach involves preprocessing chest X-ray images and their corresponding text reports using the pre-trained BLIP processor, training the novel hybrid vision-language model on paired data, and fine-tuning it for clinical relevance and coherence. The performance of CheXSBT is rigorously evaluated using established metrics such as BLEU, ROUGE, and METEOR, achieving scores of 0.232 for BLEU-4 and 0.392 for ROUGE-L, outperforming other state-of-the-art models and ensuring high-quality report generation. By reducing radiologists' workload and providing quick, accurate information, CheXSBT aims to transform the intersection between AI and clinical practice, making radiology reporting more efficient, consistent, and accessible.

Keywords: Vision-language models · Chest X-ray · Radiology report generation · Transformer · Swin transformer · BERT

1 Introduction

Radiology is a cornerstone of modern healthcare, providing essential diagnostic insights through medical imaging. However, increasing demand for imaging services, particularly in chest X-ray (CXR), has led to a growing workload for radiologists, contributing to delays in diagnosis and variability in the quality of

the report. Artificial intelligence (AI) offers a promising solution to these challenges by automating radiology report generation using vision-language models (VLMs).

A standard chest X-ray (CXR) report typically consists of three key sections: (i) the *indications* section, which outlines the referring clinician's reasons for requesting the CXR and relevant aspects of the patient's medical history; (ii) the *findings* section, which details observations derived from the radiology image; and (iii) the *impression* section, which provides the clinical diagnostic information [18]. An example of such a report is shown in Fig. 1. Among these, the indications section is particularly important as it guides the interpretation of the X-ray by highlighting the relevant clinical context. We propose an approach to automatically generate the findings section by utilising both the indications and the CXR.

FINAL REPORT

INDICATION: History: __F with malignancy, recent cycle chemo last week, DVT last month, now w/ SIRS+ presentation, malaise, JVD, epig abd pain since last night

FINDING: Herat size is normal. The mediastinal and hilar contours are unchanged, with tortuosity of the thoracic aorta again noted. Atherosclerotic calcifications are seen throughout the aorta. Pulmonary vasculature is normal. No focal consolidation, pleural effusion or pneumothorax is present. Multilevel degenerative changes are seen in the thoracic spine. Clips in the right upper quadrant of the abdomen are re- demonstrated.

IMPRESSION: No acute cardiopulmonary abnormality.

Fig. 1. An example of a Chest X-Ray and corresponding report from the MIMIC-CXR dataset.

In this work, we present a novel hybrid vision-language model called as **CheXSBT**, which is designed to automate the generation of chest X-ray (CXR) reports. The model is built by combining our custom-designed **Dual-Attention Swin Transformer (DAST)** for vision processing and **BERT** (Bidirectional Encoder Representations from Transformers) [2] for natural language understanding, creating a hybrid vision-language framework that generates structured, context-aware, and clinically relevant reports. This approach not only improves accuracy but also enhances the interpretability of the generated findings.

Our model is designed to process multimodal data by integrating visual features from chest X-ray images with textual embeddings from corresponding reports. This fusion of modalities enhances the generation of more informative and coherent reports, closely aligning with expert-written radiology interpretations.

The model is trained on the MIMIC-CXR [4–6] dataset, which comprises over 370,000 chest X-ray images paired with corresponding radiology reports. We evaluate the performance of CheXSBT using well-established natural language generation (NLG) metrics such as BLEU [15], ROUGE [8], and METEOR [1] to ensure both linguistic quality and diagnostic accuracy. Our results demonstrate that the integration of visual and textual representations significantly enhances the coherence and reliability of generated reports compared to unimodal approaches.

The main contributions of our work are:

- We present CheXSBT, a novel hybrid vision-language model that integrates our custom-designed Dual-Attention Swin Transformer (DAST) for visual feature extraction with a BERT-based encoder for language processing, specifically tailored for radiology report generation. To the best of our knowledge, this is the first work to combine DAST with BERT for automated chest X-ray report generation.
- Our custom-designed Dual-Attention Swin Transformer (DAST) employs a two-stage structure in each block, incorporating Window Multi-head Self-Attention (W-MSA) followed by Shifted Window Multi-head Self-Attention (SW-MSA), which enables the model to capture both local and global dependencies. Additionally, Layer Normalisation (LN) is applied before each attention operation to stabilise training and improve convergence. A residual connection is also included to preserve gradient flow and prevent vanishing gradients during deep network training. These modifications improve the effectiveness of the Swin Transformer in visual tasks, contributing to better performance in feature extraction.
- Our model learns to interpret medical images effectively and generate structured, clinically meaningful textual descriptions, improving diagnostic consistency. We conduct rigorous evaluations using well-established NLG metrics, demonstrating the effectiveness of CheXSBT in generating high-quality, diagnostically valuable reports.

2 Related Work

Recent advancements in AI, particularly in vision-language modeling, have greatly enhanced automated medical report generation. Earlier approaches relied on CNN-RNN architectures, but modern research has increasingly adopted transformer based models due to their superior performance in capturing complex relationships between images and text [18]. In this section, we discuss the works related to CheXSBT.

You et al. [21] addressed the challenge of limited image-text pairs in chest X-ray (CXR) datasets by generating synthetic pairs using radiologist-designed prompts. Their method employs two contrastive losses—Image Contrastive Loss and Text Contrastive Loss—to enhance image-text retrieval and classification. Sanjeev et al. [16] introduced TiBiX, a transformer-based model with causal

attention, which integrates temporal data from prior scans to enhance report generation.

Windsor et al. [19] explored improving vision-language models (VLMs) under low-data conditions, utilising unimodal self-supervision and contrastive loss functions to enhance model generalisation in report generation. Nooralahzadeh et al. [14] proposed a multi-stage generation approach (M^2TR P), where global image concepts were first extracted and then refined into detailed, coherent radiology reports using a transformer-based sequence-to-sequence model.

Liu et al. [9] introduced a contrastive attention (CA) model that enhances the representation of abnormal regions by leveraging contrastive mechanisms. Zeiser et al. [22] proposed CheXReport, a fully transformer-based encoder-decoder framework employing Swin Transformer blocks to improve the integration of visual and textual features. Nicolson et al. [13] investigated transfer learning by initialising the encoder with a Convolutional Vision Transformer (CvT) pretrained on ImageNet-21K and the decoder with DistilGPT2, demonstrating effective performance gains. Sîrbu et al. [17] introduced GIT-CXR, an end-to-end transformer-based model for generating factually complete X-ray reports, incorporating curriculum learning to enhance model performance. Wang et al. [18] proposed a multimodal approach combining R2Gen and CvT2DistilGPT2 for automated chest X-ray report generation.

Building on these advancements, we propose a novel hybrid vision-language architecture for CXR report generation. Hybrid transformer architectures remain underexplored in this domain but have demonstrated effectiveness in various natural language processing tasks, including language modeling and machine translation. By leveraging a hybrid transformer for CXR report generation, we aim to capture complex relationships between CXR images and corresponding reports while mitigating issues such as hallucination.

3 Methodology

Our novel hybrid model, CheXSBT, employs a vision-language approach for automated radiology report generation. As shown in Fig. 2(a), the architecture comprises three key components: a Vision Encoder, a Language Encoder, and a Multimodal Fusion module. By utilising a hybrid transformer architecture—with our custom-designed Dual-Attention Swin Transformer (DAST) as the Vision Encoder and BERT as the Language Encoder—CheXSBT effectively extracts richer visual features and relationships from X-ray images, while enhancing the integration of these visual insights with textual elements.

3.1 Vision Encoder

The Vision Encoder in CheXSBT is based on the Swin Transformer [10], a state-of-the-art vision transformer architecture known for its efficiency in handling large-scale visual data. The Swin Transformer utilises a hierarchical structure with shifted windows to capture both local and global features in images [10].

Unlike traditional convolution-based models, it processes images using a patch-based approach with self-attention mechanisms, enabling better scalability and efficiency [3]. In our model, the Vision Encoder extracts high-level visual features from input chest X-ray images.

As illustrated in Fig. 2(b), we have custom-designed Dual-Attention Swin Transformer (DAST) to enhance feature extraction. Each block consists of two consecutive stages, where each stage includes a Window Multi-head Self-Attention (W-MSA) layer followed by a Shifted Window Multi-head Self-Attention (SW-MSA) layer. These attention mechanisms enable the model to capture both local and global dependencies by processing patches within fixed windows and then shifting those windows in subsequent layers.

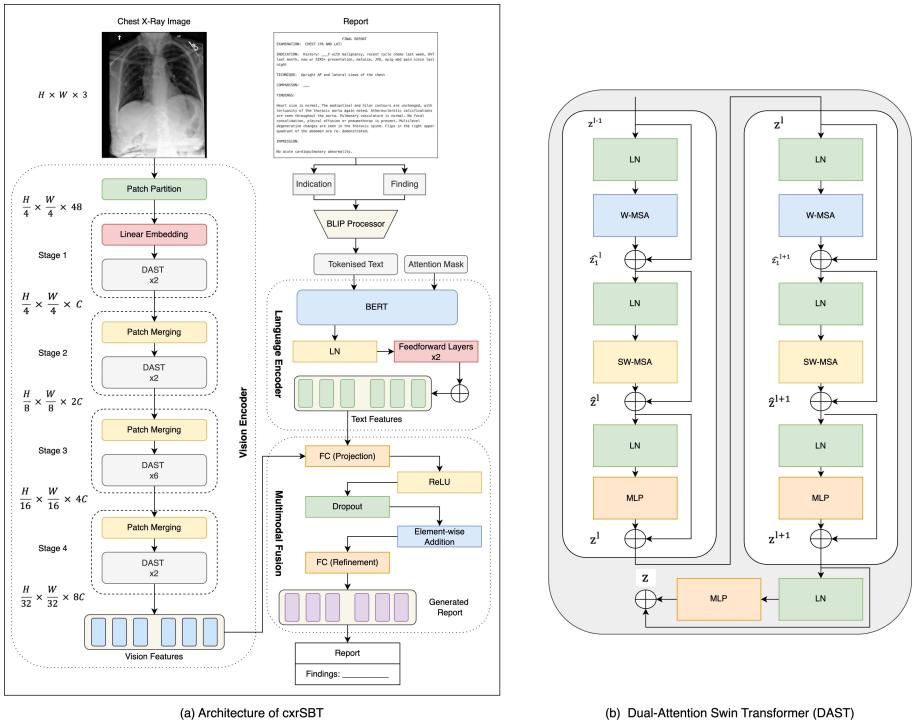


Fig. 2. (a) The detailed architecture of CheXSBT. (b) The custom-designed Dual-Attention Swin Transformer (DAST). Each block is formed by two sets of Window Multi-head Self-Attention (W–MSA) and Shifted Window Multi-head Self-Attention (SW–MSA) layers, three sets of Multi-Layer Perceptrons (MLP), and a set of Linear Normalisations (LN) before each attention and MLP operation.

The feature extraction process begins with Layer Normalisation (LN) applied before each attention operation. The output of W-MSA is computed as:

$$\hat{Z}_1^l = \text{W-MSA}(\text{LN}(Z^{l-1})) + Z^{l-1} \tag{1}$$

This is followed by the SW-MSA operation:

$$\hat{Z}^l = \text{SW-MSA}(\text{LN}(\hat{Z}_1^l)) + \hat{Z}_1^l \tag{2}$$

After the attention mechanisms, another Layer Normalisation (LN) layer is applied, followed by a Multi-Layer Perceptron (MLP) block, which refines the extracted features:

$$Z^l = \text{MLP}(\text{LN}(\hat{Z}^l)) + \hat{Z}^l \tag{3}$$

Each operation is also accompanied by a residual connection to maintain information flow. This entire process is repeated twice within a DAST Block to allow for deeper feature representation and we obtain the final equations:

$$\hat{Z}_1^{l+1} = \text{W-MSA}(\text{LN}(Z^l)) + Z^l \tag{4}$$

$$\hat{Z}^{l+1} = \text{SW-MSA}(\text{LN}(\hat{Z}_1^{l+1})) + \hat{Z}_1^{l+1} \tag{5}$$

$$Z^{l+1} = \text{MLP}(\text{LN}(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \tag{6}$$

$$Z = \text{MLP}(\text{LN}(Z^{l+1})) + Z^{l+1} \tag{7}$$

The final output of the Vision Encoder is a set of embeddings representing crucial visual features of the input image:

$$V_f \in R^{\frac{H}{32} \times \frac{W}{32} \times 8C} \tag{8}$$

These embeddings are subsequently utilised by the Multimodal Fusion component to generate radiology reports.

Comparison with Original Swin Transformer Blocks The original Swin Transformer Blocks employ a standard W-MSA and SW-MSA sequence with Layer Normalisation and MLP layers. While effective, they maintain a relatively simple design in terms of feature extraction depth and computational complexity. In contrast, our custom DAST introduce an additional MLP block at the end and an enhanced attention mechanism configuration, which allows for deeper feature abstraction and improved learning capacity.

Another key difference is in the hierarchical depth and the number of attention heads used in our design. By adjusting these factors, our modified architecture is better tailored for the fine-grained and complex features present in medical imaging, particularly chest X-rays. Additionally, our implementation optimises gradient flow and convergence stability by refining the residual connections and Layer Normalisation placements, making it more robust for training on large-scale medical datasets.

3.2 Language Encoder

The Language Encoder in CheXSBT is based on BERT-based architecture, a well-established framework for natural language understanding and generation. BERT (Bidirectional Encoder Representations from Transformers) [2] utilises a transformer-based encoder to learn rich contextual representations by considering both left and right contexts of a given token. In our model, the Language Encoder is responsible for generating textual descriptions in conjunction with the visual features extracted by the Vision Encoder.

The language encoder processes two primary inputs: text features (i.e., textual embeddings from medical reports) and attention masks, which guide the model to focus on relevant portions of the text while ignoring padding tokens. The architecture consists of multiple layers of self-attention, feedforward neural networks, and residual connections, which collectively enable the model to learn robust textual representations. During training, the language encoder is optimised to predict the next token in the sequence using both textual and visual context, ensuring that the generated text remains clinically relevant and coherent.

Given a sequence of tokenised input text X_t, the encoder outputs contextualised embeddings:

$$T_f = \text{Encoder}(X_t) \in R^{n \times d} \tag{9}$$

where n is the number of tokens and d is the hidden dimension. These embeddings capture the semantic structure of the input text and are used for downstream fusion with visual features.

To improve generalisation and reduce overfitting, the language encoder incorporates layer normalisation, dropout, and feedforward layers. By fine-tuning the model on domain-specific radiology texts, the Language Encoder effectively aligns textual features with visual cues, supporting the generation of accurate and contextually meaningful radiology reports.

3.3 Multimodal Fusion

The Multimodal Fusion Network is central to integrating the visual and textual modalities in our framework. After extracting high-level visual features via the Vision Encoder and contextualised textual features via the Language Encoder, the fusion network combines these representations into a unified format suitable for report generation.

In our model, the Multimodal Fusion Network first projects both, the visual and textual embeddings into a shared representation space using a fully connected layer, followed by a ReLU activation and dropout for regularisation. An element-wise addition is then applied to align the visual and textual information effectively. Although simple, this fusion method enables the model to leverage complementary cues from both modalities. The fused representation is

then refined through another fully connected layer, consisting of linear projections followed by non-linear activation, to enhance the integration of multimodal information and produces the final output—the generated report.

Let $V_f \in R^{\frac{H}{32} \times \frac{W}{32} \times 8C}$ be the vision features and $T_f \in R^{n \times d}$ the text features, the fused representation is computed as:

$$M_f = V_f + T_f \tag{10}$$

where M_f represents the combined multimodal features, used to generate the final radiology report.

The Multimodal Fusion Network is designed to balance contributions from both modalities, ensuring that neither dominates the final prediction. By learning a joint representation space, the model effectively captures complex relationships between images and text, improving its ability to generate accurate and informative outputs in multimodal tasks.

4 Experiments

4.1 Dataset

Our study leverages the MIMIC-CXR [4–6] dataset for training, evaluating, and testing. The dataset is publicly available and contains 377,110 chest X-ray images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA, USA. We have structured the dataset as a CSV file, which links image file paths to their corresponding reports. To facilitate training, we implement a custom dataset class that reads the CSV file, loads images from the specified directory, and applies necessary preprocessing steps. We used the MIMIC-CXR dataset's predefined split for training, validation and testing.

4.2 Data Preprocessing

To ensure compatibility with deep learning models, the preprocessing pipeline involves handling image preprocessing, text preprocessing, and data alignment.

Image Preprocessing Chest X-ray images are resized to 224x224 pixels to maintain uniform input dimensions. Pixel intensity values are normalised using standard mean and deviation values to ensure stability in training. To improve model generalisation and robustness, data augmentation techniques such as random rotations, scaling, and contrast adjustments are applied.

Text Preprocessing The Radiology reports undergo standardisation to reduce variability in medical terminology. Tokenisation is performed using the pre-trained BLIP [7] processor from Hugging Face, which converts text into structured sequences suitable for transformer-based models. Additionally, noise removal is applied to eliminate redundant or irrelevant information, ensuring high-quality textual inputs.

Data Alignment The dataset undergoes verification to maintain correct image-text pairings, ensuring consistency between the chest X-ray images and their corresponding reports. Any missing or corrupted data entries are filtered out to maintain data integrity and improve training performance.

4.3 Implementation

Our model, CheXSBT, is trained on the MIMIC-CXR dataset to generate radiology reports from chest X-ray images. For the vision encoder, we use our custom-designed Dual-Attention Swin Transformer (DAST), while for the language encoder, we employ a BERT-based architecture.

For model optimisation, we employed the AdamW [11] optimiser with an initial learning rate of $1e^{-3}$ and a weight decay of $1e^{-3}$. A StepLR scheduler is applied to decrease the learning rate by 15% after each epoch, ensuring stable convergence. The model is trained with a batch size of 8, balancing computational feasibility and performance. We perform five runs of the model with same training, validation, and test splits to ensure consistency in results.

During training, we leverage a multi-GPU setup when available, which allows for more efficient processing of larger batches. The training loop involves feeding the model pre-processed chest X-ray images along with tokenised indication texts. Predictions are generated, and the loss is computed using the cross-entropy function. Validation is performed at the end of each epoch to track performance trends and adjust hyper parameters accordingly.

We trained CheXSBT on Aire HPC at the University of Leeds, which consists of 28 GPU nodes, each containing 3 NVIDIA L40S GPUs, offering a total of 84 GPUs. For our experiments, we used a single GPU node with 8 CPUs per task and 24 GB memory per CPU. Training each epoch required approximately 8 hours. Our implementation was built using PyTorch, ensuring a robust and scalable framework for generating chest X-ray reports.

4.4 Evaluation Metrics

To quantitatively evaluate the performance of our model, we used several established Natural Language Generation (NLG) metrics such as BLEU (Bilingual Evaluation Understudy) [15], METEOR (Metric for Evaluating Translation with Explicit Ordering) [1], and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) [8].

5 Results and Discussion

In this section, we provide a comprehensive analysis of CheXSBT's performance. We begin with an ablation study to evaluate the impact of different architectural components on the model's performance. This is followed by a comparison with state-of-the-art models. Finally, we present a qualitative analysis, including both visual and textual comparisons, to showcase the model's capabilities

in generating chest X-ray reports. This in-depth evaluation aims to validate the effectiveness of CheXSBT in producing accurate and informative chest X-ray reports.

5.1 Ablation Study

We conducted an ablation study to evaluate the performance of the model by utilising different architectural components. The study began with training a pre-trained BLIP model [7] from Hugging Face on the MIMIC-CXR dataset. We then evaluated the original Swin Transformer with BERT, followed by our custom-designed Dual-Attention Swin Transformer (DAST) integrated with BERT. All models are trained with a batch size of 8, balancing computational feasibility and performance. We perform five runs of the model with same training, validation, and test splits to ensure consistency in results, and the reported scores represent the mean ± standard deviation of all runs. Table 1 presents a summary of the scores achieved by the models across the evaluation metrics.

Table 1. Ablation study using different architectural components for CheXSBT on MIMIC-CXR dataset. To ensure consistency, the models were trained and tested five times using the same training, validation, and test splits, and the reported scores represent the mean ± standard deviation of all runs. Higher values in bold denote the best results in all columns.

Model	Pre-trained BLIP	Swin Transformer + BERT	CheXSBT DAST + BERT
BLEU 1	0.323 ± 0.12	0.408 ± 0.06	**0.502 ± 0.06**
BLEU 2	0.265 ± 0.10	0.319 ± 0.08	**0.397 ± 0.06**
BLEU 3	0.198 ± 0.13	0.256 ± 0.05	**0.306 ± 0.04**
BLEU 4	0.154 ± 0.08	0.198 ± 0.05	**0.232 ± 0.03**
ROUGE-L	0.257 ± 0.09	0.301 ± 0.07	**0.392 ± 0.06**
METEOR	0.166 ± 0.11	0.199 ± 0.09	**0.229 ± 0.03**

The results from the ablation study indicate that our model, CheXSBT with the custom-designed Dual-Attention Swin Transformer (DAST) integrated with BERT, outperforms the other models across all evaluation metrics. This suggests that our model is particularly effective at generating longer, more detailed captions that accurately capture the essence of the input image. The improved performance, as reflected in the higher scores across BLEU [15], ROUGE-L [8], and METEOR [1] metrics, demonstrates that the incorporation of DAST with BERT significantly enhances the quality of medical image captions.

5.2 Comparison with State-Of-The-Art Models

We compare our CheXSBT with several other state-of-the-art models for generating chest X-ray reports on the MIMIC-CXR dataset, including Contrastive Attention (CA) [9], CheXReport [22], Convolutional Vision Transformer with the

Distilled Generative Pre-trained Transformer 2 (CvT2DistilGPT2) [13], Meshed-Memory Transformer (M^2 TR P) [14], GIT-CXR (MV+C+CL) [17], Auxiliary Signal Guidance and Memory-Driven(ASGMD) [20] and Automated Generation of Accurate & Fluent Medical X-ray Reports with Multi-view (MV), w/ clinical text (T), and interpreter (I) (AGAFMXR (MV+T+I)) [12].

Table 2. Comparison with state-of-the-art models on MIMIC-CXR dataset. All metrics for the state-of-the-art models are directed cited from the original paper. Higher values in bold denote the best results in all columns. The Δ indicates that the model was trained and tested five times using the same training, validation, and test splits, and the reported scores represent the mean ± standard deviation of all runs.

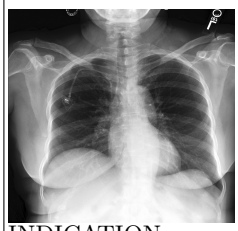
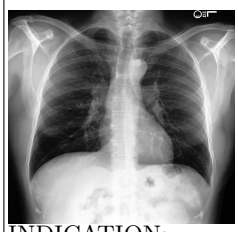
Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L	METEOR
CA [9]	0.350	0.219	0.152	0.109	0.283	0.151
CheXReport [22]	0.354	0.225	0.145	0.127	0.286	0.147
CvT2DistilGPT2 [13]	0.392	0.247	0.171	0.126	0.286	0.154
M^2 TR P [14]	0.378	0.232	0.154	0.107	0.272	0.145
GIT-CXR (MV+C+CL) [17]	0.403	0.286	0.215	0.168	-	**0.369**
ASGMD [20]	0.372	0.233	0.154	0.112	0.286	0.152
AGAFMXR (MV+T+I) [12]	0.495	0.360	0.278	0.224	0.390	0.222
CheXSBT DAST + BERT Δ	**0.502 ± 0.06**	**0.397 ± 0.06**	**0.306 ± 0.04**	**0.232 ± 0.03**	**0.392 ± 0.06**	0.229 ± 0.03

Table 2 presents the performance metrics reported by previous studies on the MIMIC-CXR dataset. The results demonstrate that CheXSBT achieves state-of-the-art performance across multiple metrics. Notably, CheXSBT outperforms existing models in BLEU-1 (0.502), BLEU-2 (0.397), BLEU-3 (0.306), BLEU-4 (0.232), and ROUGE (0.392), while also achieving competitive results in METEOR (0.229).

Despite its strong overall performance, CheXSBT exhibits slightly lower scores in METEOR compared to GIT-CXR (MV+C+CL) [17] (0.229 vs. 0.369). However, its superior performance in BLEU-4 and ROUGE suggests that CheXSBT excels in generating longer, more coherent, and contextually rich reports. This capability is crucial in medical report generation, where detailed and accurate descriptions are paramount.

The ability of CheXSBT to generate longer, more informative reports makes it particularly well-suited for clinical applications, where clarity and comprehensiveness are critical. While some state-of-the-art models prioritise shorter, more structured outputs, CheXSBT's proficiency in capturing long-term dependencies through Transformer-based architectures aligns well with the requirements of medical documentation. Its effectiveness in aligning generated reports with

Table 3. Comparison of ground truth with the results generated by the original Swin Transformer with BERT and CheXSBT DAST with BERT models on randomly selected chest X-ray images from the MIMIC-CXR dataset. The first case demonstrates a strong alignment with the ground truth, while the second case fails to accurately replicate the reference report, highlighting the possible variability in model performance.

Chest X-ray Image	Ground Truth	Swin Transformer + BERT	CheXSBT DAST + BERT
 INDICATION: ___ year old woman with lymphoma, no blood return from her port // port misplacement	The lungs are clear. The cardiomediastinal silhouette, hilar contours, pleural surfaces are normal. The tip of the right Port-A-Cath is in the upper SVC. No pneumothorax, pulmonary edema, pneumonia, or pleural effusions. Right tracheal deviation is due to enlargement of the left thyroid gland.	Lungs are clear with no signs of pneumothorax, pleural effusion, or pulmonary edema. The cardiomediastinal silhouette and hilar contours appear normal. The right Port-A-Cath tip is positioned in the upper SVC. Rightward tracheal deviation noted.	The lungs are clear with no evidence of pneumothorax, pleural effusion, pulmonary edema, or pneumonia. The cardiomediastinal silhouette, hilar contours, and pleural surfaces appear normal. The right Port-A-Cath tip is correctly positioned in the upper SVC. Mild rightward tracheal deviation is observed, likely due to an enlarged left thyroid gland.
 INDICATION: ___ year old man with 2.5 week history of cough.	Lungs are well expanded and clear. No pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are unremarkable. Hilar contours are normal. No acute cardiopulmonary process.	The lung fields are underinflated with basal opacities. Moderate left-sided pneumothorax is suspected. Cardiomediastinal silhouette is distorted. Findings are concerning for possible pulmonary edema.	There is evidence of mild pulmonary infiltrates. A small right pleural effusion is present. The heart appears mildly enlarged. Prominent hilar markings are noted. These findings may indicate an early cardiopulmonary abnormality.

ground truth radiological descriptions highlights its potential for real-world deployment in automated radiology report generation.

5.3 Qualitative Analysis

In Table. 3, we present a qualitative comparison of results generated by the original Swin Transformer with BERT and our CheXSBT DAST with BERT models, using randomly selected chest X-ray images from the MIMIC-CXR dataset. For each image, we include the clinical indication provided to the model and compare the generated reports with the ground truth, which contain detailed findings from radiologists.

In the first case, the ground truth describes a patient with lymphoma and a port misplacement, with no signs of pneumothorax or other respiratory issues. Both models generate similar reports, differing slightly in phrasing but consistently identifying key features such as clear lungs, a normal cardiomediastinal silhouette, and proper Port-A-Cath placement. Our CheXSBT DAST with BERT model provides a more comprehensive description, particularly noting the mild rightward tracheal deviation likely caused by thyroid enlargement.

The second case represents a normal chest X-ray, with no pleural effusion, pneumothorax, or cardiopulmonary abnormalities. However, both models fail to replicate the ground truth accurately. Notably, the CheXSBT DAST with BERT model introduces hallucinated findings such as mild infiltrates and an enlarged cardiac silhouette—none of which are supported by the ground truth. Furthermore, the generated report lacks fluency and coherence, deviating significantly from the concise and factual style expected in clinical documentation.

Overall, our CheXSBT DAST with BERT model demonstrates the ability to produce accurate and fluent radiology reports in many cases, there are instances where the models either omit critical information or hallucinate pathological findings. These failure cases highlight the need for improved model grounding and finer control over clinical accuracy. In future work, we aim to address these limitations by incorporating pathology-aware training objectives and fine-grained error analysis frameworks.

6 Conclusion, Limitations and Future Work

In this paper, we introduced CheXSBT, a novel hybrid vision-language model that advances the automated generation of chest X-ray reports. By combining our custom-designed Dual-Attention Swin Transformer (DAST) for visual feature extraction with a BERT-based language encoder, CheXSBT provides a state-of-the-art solution that improves both the accuracy and clinical relevance of radiology reports. This integrated architecture enables the generation of coherent, contextually rich, and diagnostically meaningful text, facilitating effective communication in medical settings. Evaluated on the MIMIC-CXR dataset, CheXSBT demonstrates superior performance over existing models in key metrics such as BLEU and ROUGE, underscoring its capability to produce comprehensive and detailed reports.

Despite these promising results, our work has several limitations. First, we acknowledge the use of element-wise addition for multimodal fusion, which may oversimplify interactions between modalities. Future work will explore more sophisticated fusion strategies, such as cross-attention mechanisms. Second, while token-level metrics like BLEU and METEOR provide useful baselines, they do not fully capture the clinical correctness or contextual relevance of generated reports. We plan to adopt more advanced, semantics-aware metrics like BERTScore in future evaluations.

We also recognise the absence of pathology-specific evaluation in this study, which could provide deeper insights into clinical performance. Future iterations of CheXSBT will incorporate fine-grained, disease-specific assessments. Furthermore, although the MIMIC-CXR dataset is one of the largest and most widely used benchmark in this field—offering broad and diverse imaging data—we acknowledge the need for generalisation across datasets. We aim to extend our evaluation to other datasets such as CheXpert Plus to assess robustness and domain transferability.

Additionally, while CheXSBT shows potential for clinical integration, embedding the model within radiology workflows—such as integration with PACS systems via DICOM-SR interfaces—to support real-world deployment can be explored in the future.

In conclusion, CheXSBT represents a significant step forward in automated radiology report generation. With further methodological refinements, broader evaluation, and clinical integration, this approach holds strong potential to improve efficiency, consistency, and scalability in radiological practice.

Acknowledgments. This work was undertaken on the Aire HPC system at the University of Leeds, UK.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization. pp. 65–72 (2005)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., et al.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
5. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database (version 2.1.0). PhysioNet (2024). https://doi.org/10.13026/4jqj-jw95
6. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0

7. Li, J., Li, D., Xiong, C., Hoi, S.C.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
8. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
9. Liu, F., Yin, C., Wu, X., et al.: Contrastive attention for automatic chest x-ray report generation. arXiv preprint arXiv:2106.06965 (2021)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Nguyen, H.T., Nie, D., Badamdorj, T., et al.: Automated generation of accurate & fluent medical x-ray reports. arXiv preprint arXiv:2108.12126 (2021)
13. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
14. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Sanjeev, S., Maani, F.A., Abzhanov, A., Papineni, V.R., Almakky, I., Papież, B.W., Yaqub, M.: Tibix: Leveraging temporal information for bidirectional x-ray and report generation. In: MICCAI Workshop on Deep Generative Models. pp. 169–179. Springer (2024)
17. Sîrbu, I., Sîrbu, I.R., Bogojeska, J., Rebedea, T.: Git-cxr: End-to-end transformer for chest x-ray report generation. arXiv preprint arXiv:2501.02598 (2025)
18. Wang, C., Janjic, V., McKenna, S.: Generating chest radiology report findings using a multimodal method. In: Yap, M.H., Kendrick, C., Behera, A., Cootes, T., Zwiggelaar, R. (eds.) Medical Image Understanding and Analysis, Lecture Notes in Computer Science, vol. 14859, pp. 177–188. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-66955-2_13
19. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Vision-language modelling for radiological imaging and reports in the low data regime. arXiv preprint arXiv:2303.17644 (2023)
20. Xue, Y., Tan, Y., Tan, L., Qin, J., Xiang, X.: Generating radiology reports via auxiliary signal guidance and a memory-driven network. Expert Systems with Applications **237**, 121260 (2024)
21. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
22. Zeiser, F.A., da Costa, C.A., de Oliveira Ramos, G., Maier, A., da Rosa Righi, R.: Chexreport: A transformer-based architecture to generate chest x-ray reports suggestions. Expert Systems with Applications **255**, 124644 (2024). https://doi.org/10.1016/j.eswa.2024.124644