Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Automated Medical Image Captioning Using the BLIP Model: Enhancing Diagnostic Support with AI-Driven Language Generation

Enas Abbas Abed[1*], Taoufik Aguili[2]

[1]Department of Computer Engineering, University of Diyala, Diyala, Iraq
[2]Department of Communications System, University of Tunis El Manar, Tunisia

ARTICLE INFO

Article history:
Received December 12, 2024
Revised April 4, 2025,
Accepted April 13, 2025
Available online June 1, 2025

Keywords:
Medical Image Captioning, BLIP Model, Radiology, Artificial Intelligence, UMLS, Diagnostic Support, Transformer Models, Deep Learning, BLEU Score, METEOR Score, ROUGE Score, Clinical Applications, Image Processing, Natural Language Generation, Healthcare AI

ABSTRACT

The interpretation of medical diagnostic images is an important activity: the number of images is growing continuously, and the number of specialist radiologists is limited globally, which often results in late diagnosis and possible clinical misinformation. The paper analyzes the BLIP model, which is an automatic medical image clinical captioning model. To refine the BLIP model, a methodology was designed based on more than 81,000 radiology images with Unified Medical Language System (UMLS) identifiers, which were obtained from the ROCO (Radiology Objects in Context) dataset. A representative subset of 1,000 images was chosen to fit within computational limitations- 800 images were used in training, 100 in validation and 100 in testing, but with the preservation of representation across major imaging modalities. They trained the model on transformer-based encoder-decoder with cross-attention mechanisms. The four key contributions of this work are (1) domain-specific fine-tuning of the model to the radiological setting, (2) the use of standardized medical terminology by using UMLS concept unique identifiers, (3) integration of explainable AI with attention heatmaps and post-hoc explanations (SHAP and LIME), and (4) evaluation of performance using accepted NLP metrics. The model attained a high semantic and clinical agreement with quantitative scores of 0.7300 (BLEU-4), 0.6101 (METEOR), and 0.8405 (ROUGE). These results prompt the idea that AI-based image captioning has a considerable potential in facilitating clinical documentation and increasing the reliability of radiological assessments.

1. Introduction

In contemporary medical practice, accurate and timely diagnostics are fundamental to effective patient care. However, numerous challenges, such as misdiagnoses, delays in test results, and the complexity of interpreting radiological scans, continue to hinder clinical decision-making. Medical imaging modalities such as X-ray, CT scans, and MRI generate vast amounts of diagnostic data, requiring expert analysis by radiologists. The increasing volume of medical images, coupled with a global shortage of radiology specialists, further exacerbates delays, increasing the risk of diagnostic errors and placing additional financial strain on healthcare systems. These challenges underscore the urgent need for AI-driven solutions capable of generating precise, standardized, and interpretable image captions to support radiologists in diagnostic decision-making.

Recent advancements in deep learning and neural networks within artificial intelligence

* Corresponding author.
E-mail address: enasabbas1111@gmail.com
DOI: 10.24237/djes.2025.18215

(AI) have significantly improved medical image interpretation, enhancing both efficiency and accuracy. In particular, image captioning models, which automatically generate descriptive text from images, have shown promise in radiology applications. Among the state-of-the-art solutions, the Bootstrapping Language-Image Pre-training (BLIP) model [1] has emerged as a powerful multimodal approach, effectively integrating visual and textual processing within an encoder-decoder framework. BLIP has demonstrated remarkable performance in general-domain image-text tasks; however, its application in medical diagnostics remains underexplored. Existing image captioning models often fail to meet the rigorous demands of medical diagnostics due to:

Lack of domain-specific knowledge; General image captioning models lack medical terminology awareness, often producing ambiguous or clinically irrelevant descriptions.

Inadequate interpretability; most models do not provide explainability mechanisms, making them difficult to validate in clinical environments.

Insufficient generalizability; many existing models perform well on general image datasets but struggle with medical images due to variations in pathology presentation, imaging modalities, and dataset biases.

Limited dataset integration; current methods rarely incorporate structured medical knowledge bases, such as the Unified Medical Language System (UMLS), which standardizes medical terminology and improves caption coherence.

To address these critical limitations, our study presents a novel framework that integrates the BLIP model with a UMLS-coded radiology dataset, offering clinically meaningful, standardized caption generation for medical images. This study introduces the following key innovations:

1. We adapted the BLIP model through fine-tuning it using a large radiology dataset which exceeds 81,000 images to obtain domain-specific patterns and medical diagnostic elements necessary for clinical practice. A customized adaptation phase helps the model create captions compatible with medical standards, thus improving the accuracy and clarity of radiology reports.
2. This approach differs from previous studies since our framework uses UMLS Concept Unique Identifiers (CUIs) during training to standardize medical terminology. The method maintains consistent terminology through this procedure, which leads to better clinical trustworthiness in the generated captions for improved educational practices and automated reporting functions.
3. The XAI techniques including attention heatmaps together with SHAP and LIME post-hoc explanations serve to demonstrate the model's caption generation process to clinicians. The successful implementation of this step builds essential trust between clinicians and AI for healthcare workflow usage.
4. A full performance assessment of the model uses current natural language processing metrics which include BLEU, METEOR and ROUGE scores. The model benefits from statistical significance analysis, which uses confidence intervals and variance to build dependable results for medical usage applications.

By enhancing AI-driven image captioning capabilities, this study provides a scalable, efficient, and clinically interpretable solution for medical diagnostics. The proposed framework bridges the gap between state-of-the-art AI models and real-world radiology applications, improving diagnostic efficiency, standardization of medical reporting, and the accessibility of AI-assisted decision-making tools in clinical practice. Through this research, we establish a foundational step toward integrating AI-based captioning models into modern healthcare systems, paving the way for faster, more accurate, and trustworthy medical imaging interpretations.

2. Related work

The most comprehensive source, the widely referred dermatology textbook regarding the interpretation of skin lesions [3], maintains that

the analysis of skin is very much like reading text. In the course of performing a skin examination, every skin lesion related to a particular disease is characterized in a certain way. A detailed discussion of these lesions is actually very similar to combining words to make a sentence or including different paragraphs. Physicians start with determining whether the lesion is macular, papular, nodular, and so forth, and what characteristics the lesion has (round, oval, brown, red with blurred margins, etc.); then the physicians categorize it as centralized or evenly distributed and so on, forming an entire "paragraph." Frequently, an all-encompassing account may result in obtaining the precise diagnosis. Hence, establishment of lesion type, shape, color, arrangement, margins, distribution and consistency forms the core diagnosis principles in dermatological field [2]. Realizing that the formulation of this problem is similar to that of image captioning, this study intends to propose an artificial intelligence algorithm replicating this interpretative process.

The image captioning system uses two steps of feature extraction followed by language modeling; textbook keywords identify symptom features yet require combined keywords for sign evaluation. It proves challenging to perform automatic assessments of generated image descriptions. The multimodal fusion architecture from Zhao et al. [3] signifies a groundbreaking development for image captioning through its ability to merge multiple types of input information for generating contextually rich and sophisticated captions. The approach combines the strengths of CNNs for extracting visual features and RNNs for textual features, through which it improves the model's comprehension of healthcare content. The integrated data system facilitates a detailed analysis of pictures, which proves essential for difficult medical tests. The high complexity of data inputs causes limitations through increased processing demands and overfitting potential, which results in inefficient training processes that become more difficult to execute. Multiple types of image captioning research exist according to the approach-based classification structure that scholars have established [4]. The researchers in [5] introduced an attribute-assisted teacher-critical training method to boost the captioning model's educational progress. In image captioning evaluation, sentence-level information is often overlooked. Contrastive semantic similarity learning was applied in [6] to capture sentence-level representation, with single-branch, dual-branch, and triple-branch model structures developed to capture different levels of detail. While attention-based models frequently focus on individual visual features, Wang and Gu [7] emphasized the relationships between image features, a critical aspect for generating coherent captions.

Medical images are increasingly integral to diagnostic processes. A learning-based framework for generating image captions was proposed in [8] to efficiently produce skin image reports. In medical diagnostics employing deep learning, there is often a trade-off between enhancing model performance and maintaining explainability. Barata et al. [9] sought to address this limitation in a skin cancer diagnostic system, noting that enhanced performance can diminish explainability due to syntactic complexity and lengthy sentences. The Siamese neural network used in that study did not yield satisfactory results for biomedical text similarity evaluation. Similarly, self-attention models, like those developed by Li et al. [10], offer significant improvements in handling long-range dependencies within data. Advantages of self-attention include its ability to focus selectively on different parts of the input sequence, important for medical images where relevant features may be spread across the image. This model is particularly effective in understanding complex sentence structures in lengthy medical reports, which can enhance the accuracy and relevance of generated captions. Limitations of self-attention models often involve their scalability and the quadratic increase in computational resources as sequence length increases, which can be prohibitive for large datasets commonly used in medical image processing. In [11], the semantic representation of each sentence was embedded for external semantic parser evaluation.

As the volume of medical images and reporting grows, it increasingly burdens

physicians. To alleviate this workload, the authors of [12] and [13] developed models to generate draft reports from related images. Given the differences between patient and normal images, a specialized X-ray image-captioning model was introduced in [12] using a decoder based on either a transformer or long short-term memory (LSTM). Notably, [13] highlighted the repetitive occurrence of specific medical terms across reports.

For continuity of medical terminology in the reports auto-generated, Wang et al. [13] developed a model which consists of template matching and sentence synthesis. To address human-centric and remote sensing image captioning, Yang et al. [14] proposed the HCCM to generate captions about human actions from associated images. Furthermore, Wang and Zhang [15] and Ye et al. [16] introduced a Visual Alignment Attention (VAA) model and a Joint-Training Two-Stage (JTTS) approach for remote sensing image captioning.

Table 1 provides a comprehensive summary of related work in the fields of image captioning and medical image analysis. It highlights the key focus areas, methodologies, problems addressed, main contributions, and datasets used by each referenced study. This comparative overview helps position the current work within the broader research landscape and identifies existing gaps and innovations in the domain.

Table 1: Summary of related work in image captioning and medical image analysis

Reference	Focus Area	Model/Method	Problem Addressed	Key Contribution	Dataset Used (if applicable)
Wolff et al., 2012 [2]	Dermatological Diagnostic Techniques	Textual Analysis in Dermatology	Lack of structured approach to skin lesion diagnosis	Describes skin lesion diagnosis as a textual composition process, requiring analysis of shape, color, and margins.	Not applicable
Zhao et al., 2019 [3]	Image Captioning Technologies	Multimodal Fusion Architecture	Difficulty in generating accurate image-text representations	Proposed a multi-modal architecture for generating image captions, integrating CNN, feature representation, language CNN, and RNN.	MS COCO, ImageNet
Bai et al., 2018 [4]	Classification Techniques in Captioning	Captioning Classification Models	Lack of standardized classification in captioning approaches	Categorized different image captioning techniques, providing a taxonomy for model comparison.	Various open-source datasets
Huang et al., 2022 [5]	Training Strategies for Caption Models	Attribute-Assisted Learning	Ineffective learning strategies in caption models	Introduced an attribute-assisted teacher-critical training strategy to enhance model learning efficiency.	MS COCO
Zeng et al., 2022 [6]	Semantic Analysis in Captioning	Contrastive Learning Models	Loss of sentence-level semantic coherence	Applied contrastive learning to improve sentence-level semantic similarity in captions.	Private dataset
Wang et al., 2022 [7]	Visual Attention Mechanisms	Visual Relationships Focus	Limited attention mechanisms in image captioning	Focused on modeling visual relationships to improve contextual	MS COCO

				coherence in generated captions.	
Wu et al., 2022 [8]	Automated Reporting for Dermatology	Learning Frameworks for Skin Imaging	Need for automation in dermatology report generation	Developed a learning-based framework to automatically generate structured skin image reports.	ISIC (International Skin Imaging Collaboration)
Barata et al., 2021 [9]	Challenges in Diagnostic Explainability	Siamese Networks for Skin Cancer	Trade-off between model performance and explainability	Addressed challenges in explainability for skin cancer diagnostics, noting the impact of syntactic complexity.	HAM10000 dataset
Li et al., 2021 [10]	Handling Complex Syntax in Medical Texts	Self-Attention Models	Difficulty in processing complex medical sentences	Designed a self-attention model to enhance readability and structure in lengthy medical reports.	MIMIC-CXR
Bölücü et al., 2023 [11]	Sentence-Level Semantic Evaluation	Semantic Embedding Techniques	Lack of semantic representation in NLP captioning	Used sentence-level embeddings to improve the external evaluation of semantic parsers.	Not specified
Park et al., 2021 [12]	X-ray Image Captioning	Transformer/LSTM Decoders	Need for specialized medical image captioning	Proposed an X-ray captioning model using Transformer or LSTM decoders to improve diagnostic descriptions.	MIMIC-CXR, IU X-ray
Wang et al., 2022 [13]	Medical Terminology Consistency	Template Matching and Sentence Synthesis	Inconsistency in medical terminology usage	Developed a unified model to ensure consistent terminology in AI-generated medical reports.	CheXpert
Yang et al., 2022 [14]	Human-Centric Image Captioning	Human-Centric Captioning Model (HCCM)	Lack of human-action descriptions in image captioning	Created a model specifically designed to generate captions for human activities in images.	MPII Human Pose Dataset
Zhang et al., 2019, Ye et al., 2022 [15, 16]	Remote Sensing Captioning	Visual Alignment and Joint-Training Models	Difficulty in aligning visual and textual features for remote sensing	Proposed visual alignment and joint-training methods to enhance remote sensing image captioning.	Remote Sensing Image Dataset

3. Proposed methodology

The schema depicted in Figure 1 shows a step-by-step process for BLIP (Bootstrapping Language-Image Pretraining) model-based automated medical image captioning that produces descriptive outcomes relevant for clinical use. ROCO dataset serves as the foundational component of this workflow because it contains a large array of radiology images that consist of X-rays and CT scans as well as MRIs and angiography. The model

requires standardized data from medical images so preprocessing ensures proper data preparation during the process of format transitioning and resolution standardization. Preprocessing operates on the data by performing manipulations, normalization processes, splitting the dataset and resizing or scaling images. To facilitate effective model training and evaluation the dataset needs to be split into training, validation and test sets while all images must undergo preprocessing to achieve consistent 256 × 256 pixels resolution and standardized format.

The processed dataset then proceeds to be evaluated using the BLIP model featuring an encoder-decoder framework based on transformer operations. Through its feed-forward layers and self-attention processes the image encoder retrieves important visual characteristics to detect complex medical patterns as well as anatomical structures together with pathological markers. The pre-processed features move through three essential modules named ITC (Image-Text Contrastive Learning) and ITM (Image-Text Matching) and LM (Language Model). The ITC module connects visual image features to textual content in order for the model to develop associations between picture patterns and related medical vocabulary. The ITM module builds upon this system because it evaluates caption relevance through analysis of text descriptions in the dataset to strengthen semantic ties. The LM module (Language Model) applies cross-attention and causal self-attention layers to produce the last captions that maintain both clinical meaning and grammatical structure and contextual coherence.

The NLP evaluation process uses three common metrics known as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) for evaluating generated captions. A BLEU score computes n-gram matches between automated text and normal reference examples to determine syntactical agreement between them. Because the METEOR score manages synonyms and paraphrasing it proves highly valuable for measuring medical text consistency. The ROUGE evaluation method analyzes complete sequences to verify that the produced captions include all vital diagnostic terminology present in baseline descriptions. The results' robustness can be validated through the application of statistical methods which determine stability and reliability of generated captions within multiple system runs.

A clinical output from this pipeline generates precise medical image descriptions which enable rapid standard radiology reporting. An automated captioning system presents multiple benefits to radiologists through streamlined workflows and quicker diagnosis and standardized medical documentation. Radiologists can utilize attention heatmaps together with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to examine how the model develops captions as well as to prove its accuracy for clinical use. The future development will center on data diversity expansion as well as model optimization for complicated multi-label captions alongside on-site healthcare professional verification to achieve real-world medical application integration.

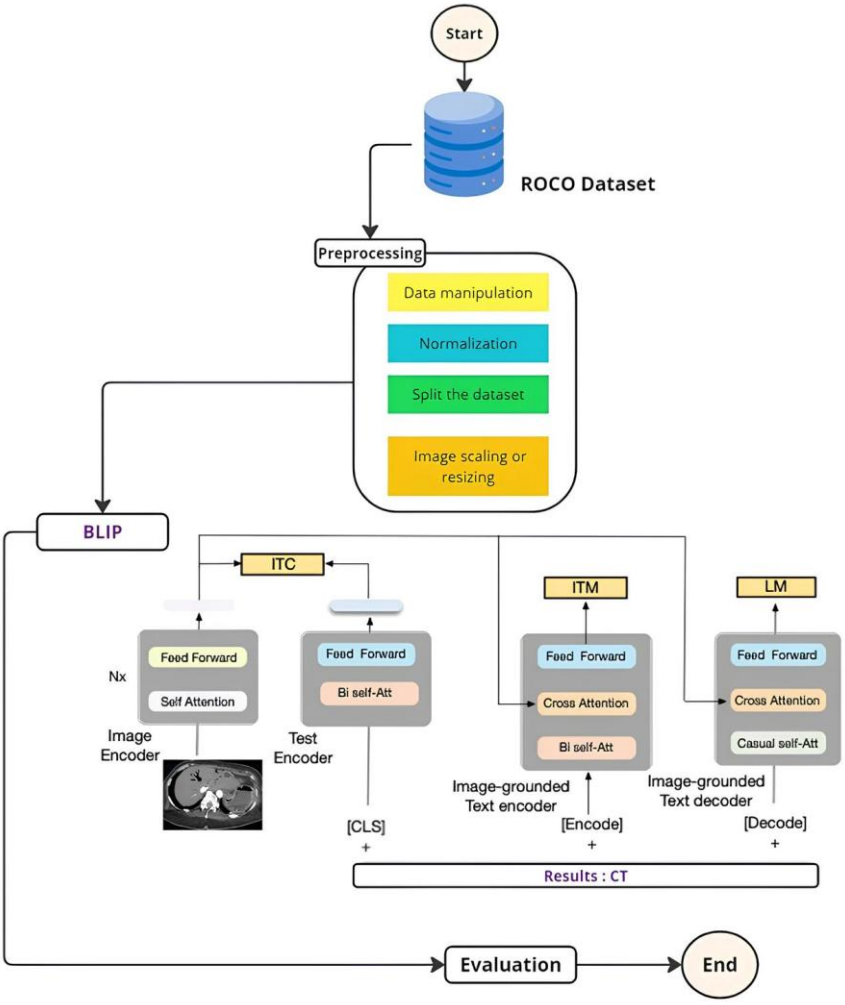


Figure 1. Proposed scheme

3.1 Data overview

This research analysis deploys more than 81,000 medical images which were obtained from different imaging modalities. This database applies different imaging types like chest X-rays as well as computed tomography (CT) scans and magnetic resonance imaging (MRI) scans and angiography images which provide universal use throughout multiple diagnostic applications. The dataset included medical image captioning pairs which combine detailed written text that describes significant clinical observations with each single medical picture. Multiple imaging formats combined with structured annotation types give the database exceptionally strong capabilities for AI instruments to create medical reports and establish precise diagnoses.

Table 2 presents a detailed summary of the dataset used in this study, outlining key attributes such as the number of images, imaging modalities, caption statistics, UMLS integration, preprocessing and augmentation techniques, as well as the dataset split for training, validation, and testing. This overview provides essential context for understanding the dataset's structure and suitability for training image captioning models in the medical domain.

The radiology images in this paper were obtained by the ROCO (Radiology Objects in COntext) dataset: a large publicly available dataset, which comprises more than 81,000 annotated radiology images published in the literatures. ROCO contains Unified Medical Language System (UMLS) concept identifiers, which provide semantic stability of medical terms. The data can be found in [28].

The dataset's images come with at least one medical description annotation which provides extensive medical content. The dataset contains approximately 2.5 million image captions, which contributes to various description methods that enhance training efficacy. The captions remain between 8 and 25 words and they measure an average of 15 words each while maintaining detailed information yet staying direct. The dataset features an essential feature that employs Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) to standardize medical vocabulary present in different reports. The inclusion of CUIs into the dataset establishes standardized medical diagnostic terms while making the data valuable for using clinical AI techniques. According to the dataset, a single X-ray image gets two UMLSCUIs (C3541877 and C0817096) when it includes the diagnostic term "dextrocardia." This ensures standardization across medical classification and diagnosis systems.

The deep learning applications required high-quality images from the dataset, so researchers performed extensive processing on all inputs. Deep learning models received standardized (224 × 224 × 3) dimension images throughout the preprocessing process. The team employed noise reduction and contrast enhancement methods primarily on CT and MRI images so that imaging artifacts would decrease while image clarity would increase. Performance achieved stability during training by using data augmentation procedures which included rotating images along with both horizontal and vertical flippings and adjusting brightness levels to create more generalized outputs.

The dataset was thoroughly split in three subsets to maximize training performance and reliability of evaluation, regarding the limitations on computational resources. To balance between the modality and diversity of captions and at the same time keep the resource usage manageable, a representative sub-set of 1,000 images was sampled out of the complete ROCO dataset :

Training Set: 800 images for model learning.

Validation Set: 100 images for hyperparameter tuning and performance optimization.

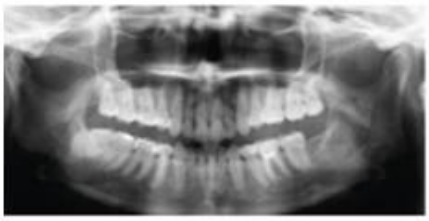
Test Set: 100 images for final model evaluation and unbiased performance assessment.

With diverse imaging modalities, well-organized medical language, and high-quality annotations, this dataset is considered a benchmark dataset for developing AI-based medical imaging studies. This focus on integrating UMLS CUIs, modality diversity, preprocessing enhancements enables the support for automated medical image captioning, thereby improving the diagnostic efficiency and consistency for various clinical applications.

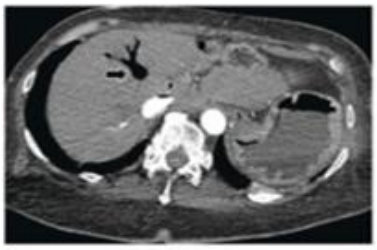
Table 2: Dataset summary table

Attribute	Details
Total Images	81,000+
Imaging Modalities	X-rays, CT scans, MRI scans, Angiography
Number of Captions per Image	2.5 (on average)
Average Caption Length	8–25 words (mean: 15 words)
UMLS Intégration	Yes (Concept Unique Identifiers for standardization)
Dataset Preprocessing	Resized (224 × 224 × 3), noise reduction, contrast enhancement
Data Augmentation	Rotation, flipping, brightness adjustment
Training Set Size	800 images
Validation Set Size	100 images
Test Set Size	100 images

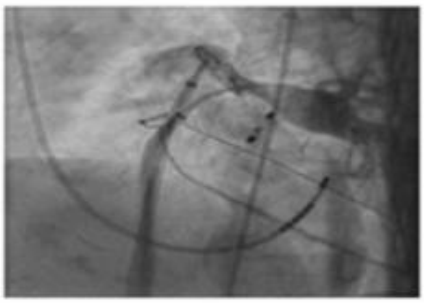
ROCO Captions	Chest X-ray showed dextrocardia
ROCO UMLS	C3541877, C0817096
Our modality label	X-ray



ROCO Captions	Panoramic X-ray taken four years later showing a unilocular radiolucent area in the left ramus
ROCO UMLS	C1561543
Our modality label	X-ray



ROCO Captions	CT image of the liver shows air in the hepatic veins (arrow) and air-contrast level in the inferior vena cava (arrowhead)
ROCO UMLS	C0023884, C1696103, C0042449, C0205054
Our modality label	CT



ROCO Captions	Angiography of LIPV shows the compression and narrowing of ostium
ROCO UMLS	C0021153, C0002978, C1829459
Our modality label	Angiography

Figure 2. Sample of the dataset

Figure 2 presents a sample from the ROCO dataset, showcasing the diversity of medical imaging and the standardized captions associated with each image.

Imaging modality variation and supply of standardized medical terminology, which make the model capable of addressing multiple image-captioning tasks. This, in turn, increases the model's ability to produce a better-quality set of captions for clinical applications, which are crucial in medical diagnosis and reporting. These are some pictures from the ROCO dataset: Only text descriptions are presented as captions, but UMLS identifiers and modalities are also pointed out. One picture is of the typical chest X-ray, and the detail at the bottom of the picture has the word "dextrocardia" written on it. This is alongside the UMLS codes (C3541877, C0817096) that refer to the particular medical concepts and the 'Modality' label that labels the image as an 'X-ray.'

The second interpretation regards it as a panoramic X-ray, illustrated with a caption which states that an image has been taken to show a radiolucent area in the left ramus. This description is linked with the code C1561543; the modality label reasserts that it is an "X-ray." The third image is the abdominopelvic CT scan that gave the caption that states that there is air in the hepatic veins and an air contrast level in the inferior vena cava. Other related medical terms use UMLS codes C0023884, C1696103,

C0042449, C0205054, while the label under the modality categorizes the image as a CT.

The last picture in the figure is an angiography image, and the description of the image explains that the LIPV (left inferior pulmonary vein) is compressed and narrowed at the ostium. Consistent UMLS codes are C0021153, C0002978, C1829459; this image gets the modality label.

These examples illustrate the heterogeneity of imaging studies contained in the ROCO dataset and illustrate how each image is associated with the uniformly generated captions and UMLS codes. This approach is useful to bring context to each image and, furthermore, to maintain medical language uniformity across the dataset, which is crucial in training models to generate accurate image descriptions.

This investigation segmented its data into three subsections for training purposes and validation tests and subsequent testing procedures. The training subset contained 800 data sets that used (224, 224, 3) for image size alongside 145 features and labels. The validation set utilized 100 samples structured in the same manner as training samples, with matching (224, 224, 3) image dimensions across all subsets. The test samples included one hundred examples with analogue data organization to both training and validation data sets. The data arrangement through partitioning allowed model developers to perform training and hyperparameter optimization and testing while preserving data integrity and achieving generalized results. The researchers chose (224, 224, 3) as the image dimension, which matches state-of-the-art deep learning models' requirements for processing visual information across all subsets.

3.1.1 Dataset sampling strategy

The ROCO (Radiology Objects in Context) dataset ROCO (Radiology Objects in Context) is a collection of more than 81,000 annotated radiology images of various modalities (e.g., X-rays, CT scans, MRIs, and angiography) with captions and UMLS identifiers. Because of the computational resources limitation, this research empirically sampled 1,000 images to make sure the training and evaluation are affordable, and the selected images maintain the modality and semantic diversity of the dataset. This subset was further partitioned into 800 training images, 100 validation images and 100 test images. The selection process was such that it offered a balanced selection of medical findings, modalities and caption complexity. Such an approach allowed effective experimentation without sacrificing the performance or generalizability of the models, as is evidenced by the high evaluation scores.

3.2 Preprocessing

Before training deep learning models for caption generation of radiology images, dataset preprocessing serves as a vital step to standardize and balance data for efficient model training purposes. Using the entire dataset of more than 81,000 medical images from various modalities including chest X-rays, CT scans, MRI scans, and angiography was unfeasible for computational processing. The research team selected a subset of 800 training samples as well as 100 validation samples and 100 test samples to address resource restrictions without sacrificing dataset variety. The chosen portion maintains equal numbers of normal and abnormal medical images for various imaging techniques, allowing the model to understand various diseases.

The specialist team applied a uniform 256 × 256 pixels resolution for all imaging files to guarantee equal processing while maintaining clinically important image features. Smooth processing becomes possible in the BLIP model’s ViT encoder because all inputs follow standardized dimensions. Class imbalance presents a major problem for medical imaging datasets because normal chest abnormalities dominate over rare disease cases. A combination of data augmentation methods was used especially for minority classes through image adjustments including rotation and flipping and contrast changes and brightness manipulations. The additions boost dataset variability through artificial means, which improves the model's capacity to handle new

unseen data points without falling into dominant class overfitting.

The presence of noise and imaging quality inconsistencies resulting from equipment inconsistency as well as patient-related factors affects radiology images. The application of Gaussian filtering for noise reduction and the implementation of the histogram equalization technique for improved contrast accomplished better visibility of diagnostic elements. The preprocessing operation protects the model from corrupting meaningful visual features by discarding unnecessary imaging noise, thus enhancing its eventual captioning accuracy.

The BLIP model needs image and text processing together, so caption preprocessing worked as a vital aspect for the model. Each caption received processing from the Long Range Arena (LRA) Bootstrapping Language-Image Pretraining (BLIP) tokenizer that converted the written descriptions into numerical tokenized sequences. Sequence padding was implemented to normalize the inputs by extending shorter captions to match the length of the longest captions. Special markers were incorporated for start and end points of captions so the decoder could identify sentences while generating logical outputs. Standardizing the medical data before entering it helps the model acquire organized clinical terminology while lowering textual errors present in the final reports.

A custom collate function optimized the computational performance by delivering proper batch preparation. The custom collate function extends all captions through dynamic padding so they match the length of the most extended sequence without any information reduction. The system generated attention masks that made it possible for the decoder to distinguish between real text tokens and padding tokens so it only dealt with meaningful words. The processed images together with tokenized captions automatically moved between computing devices (CPU/GPU) for efficient training operations.

These preprocessing methods helped develop a clean, well-balanced dataset which maintained a structured data format to prevent performance-damaging biases. The modifications boost BLIP-based medical image captioning model performance so it can produce accurate clinical descriptions from various medical imaging sources. These initiatives serve the purpose of developing AI-enabled automatic medical reporting which enhances diagnostic capability while supporting healthcare practitioners during their decision-making process.

3.3 Rationale for Choosing BLIP

The BLIP (Bootstrapping Language-Image Pretraining) model served as our choice for medical image captioning automation because its outstanding performance matched medical diagnosis requirements. BLIP stands apart from generalized text-image training such as ViLT and LXMERT because it provides domain-specialized capabilities that tailor to medical image-language pairs. The medical field depends on this capability to achieve medical terminology precision together with accurate captions in context. The vision and language pretraining capability of BLIP enables better understanding of complex image-text connections, resulting in superior performance for medical caption generation.

The Vision Transformer (ViT) architecture within BLIP performs better than the convolutional neural network (CNN)-based models such as LXMERT, thus making it suitable for vast and varied medical imaging datasets due to its better scalability and interpretability features. Standard medical terminology maintenance is possible through BLIP integration with the Unified Medical Language System (UMLS). BLIP represents an ideal solution for diagnostic efficiency improvement and radiologist decision support based on its advanced capabilities and detailed image description processing through enhanced cross-attention mechanisms.

3.4 BLIP Model

In image captioning, the BLIP (Bootstrapping Language-Image Pre-training) model is also created to be a transformer-based encoder-decoder model. For this purpose, we utilize the BLIP model and processor, and the

BLIP model used in this study is "Salesforce/blip-image-captioning-base." The BLIP model comprises an image encoder and a text decoder which have been connected through a cross-attention layer to facilitate the two for interoperation and understanding of both image and textual inputs.

Image Encoder: The input images are fed through the image encoder, where the convolutional layers and self-attention layers extract further abstract visual features. The encoder employed in this study is a Vision Transformer (ViT) where the image is first divided into patches, then linearly embedded and processed through transformer layers. Let X ∈ RH×W×C be defined to be an input image whose dimensions are represented by H, W, and C being the height, width and the number of channels, respectively. The image is then split into patches of size P × P, which produces a sequence of flattened patches Xp ∈ RN×(P²·C) where N = H·W/P², H and W are the height and width of the image respectively, and C is the Each patch is linearly projected and encoded as

$$Z_0 = \left[X_p^1 E; X_P^2 E; \ldots; X_P^N E\right] + \mathrm{E_{pos}}, \qquad (1)$$

Where E represents the embedding matrix and Epos denotes the positional embeddings.

Text Decoder: The text decoder is an RNN transformer that generates captions conditioned on two components, previously produced words and the output of the image encoder. A cross-attention mechanism is used in the decoder to focus in on the image features and, in parallel, decode the textual components to capture the context. By y<t, it is meant the sequence of tokens produced up to the previous time step t. The decoder computes the probability of the next token given the image features Z and previous tokens:

$$\mathrm{P}\left(y_t | y < t, Z\right) = Softmax\left(W_o h_t\right), \qquad (2)$$

Where W_o is the output projection matrix, and h_t is the hidden state computed by the transformer layer for time step t.

An RNN transformer structure powers the text decoder because it processes sequences quickly while maintaining word context throughout the generation process. The model shows excellent performance for medical image captioning because it considers word sequences for generating precise descriptions. Through its combination of recurrent neural networks (RNNs) alongside transformer-based self-attention operations, the decoder maintains control over distant dependencies within caption sequences because it receives information from earlier textual outputs and visual image encoder features. The design choice integrates sequential language generation capabilities with transformer-based parallel attention because it enables the system to produce detailed contextually relevant medical text.

Cross-Attention Mechanism: The cross-attention layer in the decoder enables interaction between image and text features by computing attention scores between the encoded image features Z and the current state of the text sequence. The attention scores are calculated as:

$$\text{Attention (Q,K,V)} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \qquad (3)$$

where Q, K, and V are the query, key, and value matrices derived from the image features and text embeddings, and d is the dimensionality of the key vectors. The cross-attention allows the model to align image regions with corresponding textual descriptions, improving the relevance and accuracy of generated captions.

Training Procedure: We define the training objective as a conditional generation task. Given an input image and its corresponding caption, the model is trained to minimize the negative log-likelihood of the caption sequence. Let Y = {y1, y2, . . . , yT } be the caption sequence with length T. The loss function is defined as:

$$\mathrm{L} = -\sum_{t=1}^{T} \log p\left(y_t | y < t, Z\right), \qquad (4)$$

Where $p(y_t|y<t, Z)$ is the probability of generating token y_t given the previous tokens and the encoded image features. This training procedure enables the model to learn an effective mapping from images to text.

Hyperparameters and Model Configuration: The proposed BLIP model's architecture has an encoder and decoder structure called transformer, with a hidden size of 768, 12

attention heads, and includes 12 layers. The model is trained with the help of the AdamW optimizer with a learning rate of $5 \times 10^{-}5$. Next, we train the model for 25 epochs with batch size 4, and using a custom ninordertopad. By utilizing this type of architecture and training, the BLIP model can generate captions that are word perfect because it integrates visual and textual inputs and will prove useful in the medical image captioning task.

Table 3 outlines the key parameters and corresponding values used for configuring the BLIP model in this study. This parameter summary provides insight into the model setup and training environment.

Table 3 : Model Parameters and Values for the BLIP Model.

Parameter	Value
Hidden Size	768
Number of Attention Heads	12
Number of Layers (Encoder and Decoder)	12
Patch Size	16 × 16
Embedding Dimension	768
Optimizer	AdamW
Learning Rate	5×10^{-5}
Batch Size	4
Number of Epochs	25
Image Input Size	256 × 256

BLIP serves image captioning functions across various domains, but medical professionals optimized it for medical images through specialization for medical datasets along with their detailed interpretation needs. Medical images differ from typical image captioning use cases because they contain highly specialized domain content including the appearance of anatomical features together with pathological conditions and procedural elements. This model applies the Vision Transformer (ViT) encoder because its design enables precise analysis of image patches for detecting faint visual cues. This technology delivers significant advantages to medical diagnosis due to the fact that diagnostic information rests in minimal differences between different tissue features. Medical-specific vocabulary along with UMLS concept embeddings have been added to the text decoder of this model, making it generate captions that follow medical terminology standards. The specific training data of BLIP consisting of X-rays and CT scans and MRIs and detailed medical captions enables it to process healthcare imagery most effectively. The specialized training has empowered the model to generate precise clinical descriptions of images that supply meaningful diagnostic information, which facilitates better medical choices.

3.5 Explainable AI

The adoption of Explainable Artificial Intelligence systems depends heavily on clinical acceptance because hospitals require transparent and trustworthy methods in their operational infrastructure. The study adopted attention heatmaps as the prime visualization method to show which image parts influenced the model's predictive behavior. Inference brought forward heatmaps that got displayed together with original medical images to reveal which parts primarily affected the caption production. The approach let us confirm that the model analyzed key pathological indicators while avoiding unnecessary image components. The analysis of input-feature to output-caption relationships employed post-hoc explanation tools which incorporated the SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) algorithms. Application of these techniques revealed what elements within images and text proved most significant for determining the end result. The evaluation procedure helped recognize possible bias during model decisions

and made possible adjustments to training for better conformance with clinical goals.

The implemented XAI methods boosted our model's interpretability along with creating an organized approach for doctors to validate and comprehend the AI-produced results. Our research incorporated XAI methods to create transparent, understandable and verifiable predictions from the model, which led to increased clinical model trust.

3.6 Evaluation metrics

The evaluation of the generated captions is conducted using three widely recognized metrics in natural language processing and image captioning: BLEU, METEOR, and ROUGE. All these and the following metrics evaluate the quality of the generated captions with respect to the reference captions from the dataset.

BLEU (Bilingual Evaluation Understudy): The BLEU score is employed to rate the degree of match reflected by n-gram overlapping between the generated captions and the standard captions. It quantifies the extent that the text generated resembles the reference text based on sequences of tokens, which determines the precision of the words generated as compared to the reference text. In our evaluation, we calculate BLEU scores for various n-gram levels: The BLEU-1, BLEU-2, BLEU-3 and BLEU-4 give the count of 1-gram, 2-gram, 3-gram and 4-gram overlap respectively. The BLEU score is calculated as follows:

$$\text{BLEU} = \exp\left(\min\left(1-\frac{len_{ref}}{len_{gen}},0\right)\right) \prod_{n=1}^{N} precision_n \quad (5)$$

Where len_{ref} and len_{gen} _gen are the lengths of the reference and generated captions, respectively, and precisionn represents the n-gram precision. The evaluation of the generated captions is conducted using three widely recognized metrics in natural language processing and image captioning: BLEU, METEOR, and ROUGE. rank assignment 1 focuses on how similar the generated captions are to the reference captions of the dataset, while rank assignment 2 focuses on captions similarity and rank assignment 3 measures how far away the generated captions are from the best matched reference captions.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): The METEOR score determines the similarity of the captions computed as the number of synonyms, stemmed words or matching literal words in the automatically generated text relative to those in the reference text were found. For the generation of the captions, there is a need to ensure that the evaluation gives more comprehensive results rather than the precision offered by BLEU; METEOR does this since it also works within the realm of recall. METEOR is ideal for scoring medical captions because it considers the possibility of finding variations of similar meaning in medical parlance. The METEOR score therefore ranges between zero and one and is computed from the harmonic mean of precision and recall and with an encouraging factor equal to the length of the phrases in the reference text.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Based on the recall and by computing the overlapping of n-grams, sequences and word pairs between the generated captions and the reference captions, ROUGE is calculated. In this study, we used the ROUGEL as the common regard, which is established on the longest common subsequence (LCS) with the generated as well as the reference text. The results are given by ROUGE-L, which gives an understanding of how comprehensively the generated captions can be said to have addressed the reference captions. It assesses sequence-level similarity rather than word matching. The evaluation procedure addresses the comparison of a list of captions produced with the test data to the captions list from within the dataset. The BLEU, METEOR, and ROUGE metrics are given for each caption pair, and the means of these are used as measures of the overall performance of the model. Overall, higher scores in these indicators reveal that there is a positive correlation between the generated captions and the reference captions through the BLIP model when captioning radiology images.

4. Experimental results

In this section we describe the results provided by the evaluation of generated captions based on BLEU, METEOR, and ROUGE scores. The mean of BLEU-1 read is 0.7959, which confirms a high amount of single words' repetition in the generated and reference captions. This score is calculated based on the model's rendition of individual word accuracy, implying that the garnered captions comprise many of the important key words of the referent descriptions. This strong unigram precision is also preferable in medical image captioning because precise terminology is crucial to representation. At the n-gram level greater than 2, the average BLEU-2, BLEU-3, and BLEU-4 are 0.7714, 0.7486, and 0.7300 respectively. As highlighted before, these results depict a diminishing score with the increase in the size of the n-gram, a common occurrence when dealing with natural language generation related issues. The high BLEU-2 score can confirm that the model works fine with bi-grams and the structure and the flow between two words nicely match the choices of reference captions. The BLEU-3 and BLEU-4 scores, although lower to the BLEU-1 and BLEU-2 scores, also remain significantly high, implying that the proposed model effectively learned to incorporate trigrams and four-grams, and performs the task of ensuring the coherent and logically consecutive arrangement of sentences – albeit slightly lesser proficiently over longer phrases – as it does for the individual sentences. This sequential accuracy is significant in the medical image captioning since it enables a proper order of the generated captions, with terms and their relation to one another understood clinically. The METEOR score of 0.6101 also aversed the method of building the model and the scale of the accuracy of the precision and recall measures. In contrast, METEOR does take into consideration recall as it measures how effectively the generated captions overlap the reference captions are. The fairly elevated METEOR score means that the model chooses not only the most appropriate words but also covers all the semantic fields specified in the reference captions. This balanced performance is especially important in medical applications because the method suggests that the model does not compromise accuracy for generalization and vice versa: it captures the semantic context of the descriptions, including all the possible synonyms and variations of the wording. Once more, a score of 0.8405 is relatively high in the case of longer sequences, confirming the proximity of the generated and reference captions according to the ROUGE metric. ROUGE encompasses the length of the actual overlap, which is demonstrated by the ROUGE-L specific to this metric; the degree in which the captions generated emulate the structures and sequences of the reference captions are encompassed in the measurement. These high ROUGE scores indicate that the generated captions preserve a good syntactic and semantic cohesion and commute the clinical information as was intended with little variation with the reference descriptions. Collectively, these findings show that BLIP model trained in this research yields satisfactory results to diverse evaluation criteria; when tested with METEOR, ROUGE, and BLEU, good numerical values are obtained when generating medical image captions. The learner model has been able to accomplish these scores – these scores indicates that the model is suitable for the task of medical image captioning as it is capable of addressing the precision of terms, sequencing, and overall coverage of clinical data. This performance also implies the usefulness of the model in clinical situations with the possibility of generating informative and relevant image captions for radiology use. Table 4 presents the experimental results of the model evaluated using standard captioning metric. The average scores reflect the model's effectiveness in generating coherent and semantically accurate medical image captions, demonstrating strong performance across all evaluation metrics.

Table 4: Experimental results of the model on BLEU, METEOR, and ROUGE Metrics

Metric	Average Score
BLEU-1	0.7959
BLEU-2	0.7714
BLEU-3	0.7486
BLEU-4	0.7300
METEOR	0.6101
ROUGE	0.8405

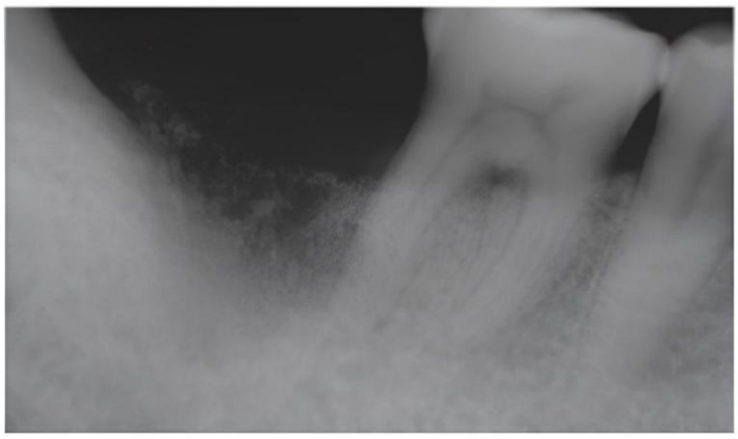


Figure 3. Generated and real captions describe periapical radiograph after autogenous socket grafting procedure accurately.

The model shows capability to produce accurate medical descriptions in Figure 3 through its handling of periapical radiograph images which serve to check root structures and adjacent bone tissue after autobt socket grafting treatment. The generated caption mirrors the reference caption when stating "periapical radiograph taken soon after autobt socket grafting." The model demonstrates excellent ability to detect crucial anatomical features ("periapical radiograph") together with detailed procedural terminology ("autobt socket grafting"), which make both components necessary for postoperative evaluations. Standardized medical terminology found in diagnostic reports achieves strong replication by the model, thus reducing the ambiguity in these reports. The slight difference between "soon after" and "immediately after" temporal expressions demonstrates a weakness, because although the model successfully understands procedural sequences, stakeholder needs further refinement to replicate exact clinical timing precision. A radiograph exclusively assays the apical part of the tooth to enable thorough assessment of both bone density and graft integration for surgical evaluation. The model accurately matches the reference caption, which indicates its capability to link visual elements such as bone texture along with graft placement with clinical action terms. The approach demonstrates efficient workflow management through a system that requires domain-specific refinement to handle detailed temporal and quantitative descriptions in future versions.

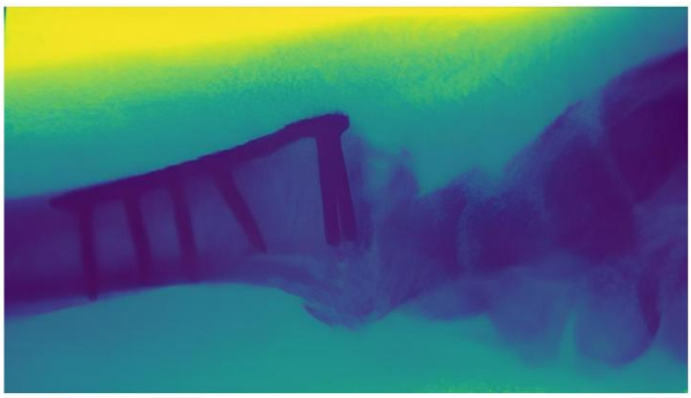


Figure 4. Model accurately identifies right subclavian CVC placement in optimal position on X-ray.

Figure 4 shows an X-ray of the subclavian region where captions confirm that a right subclavian CVC maintains an optimal position. The exact correspondence between model-produced captions and reference text proves that the model effectively recognizes relevant

clinical vocabulary and correctly detects medical device placement. The image concentrates on the right subclavian area because it shows a central venous catheter (CVC) placement for essential medical procedures in fluid delivery and medication treatment and hemodynamic observation settings. The correct placement for catheters, which meets clinical requirements, is called optimal position, because improper positioning might generate complications including thrombosis and vascular injury along with malposition. The model demonstrates aptness for functionally assisting radiological interpretations and clinical strategic decisions by producing accurate references. The model demonstrates the ability to detect relevant anatomical areas and medical equipment which could improve operational productivity in active healthcare settings, thereby reducing workloads for radiologists and medical staff. The verification procedure evaluated the model's ability through specific case analysis, but a thorough examination must assess its performance when encountering misplacements and anatomical changes or catheter kinking and migration conditions. The evaluation of this model requires identifying its reliable usage boundaries because its ability to detect minor deviations from typical catheter placements remains unclear. Future research needs to evaluate the detection capacity of this model to identify inferior catheter placements and possible complications before considering its implementation in clinical diagnosis.

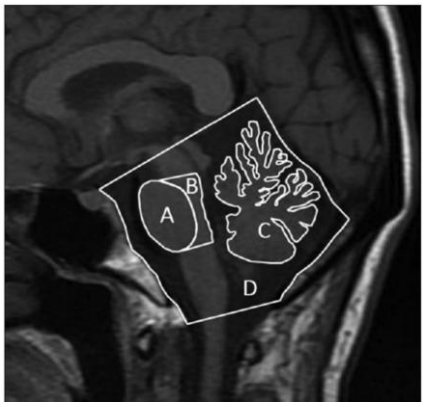


Figure 1.MRI image caption highlights small residual cavity with debris post-abscess treatment.

The brain data shows a sagittal T2-weighted turbo spin-echo (TSE) MRI scan with model-generated text along with human-produced captions in Figure 5. Head Medical professionals used a sagittal T2-weighted TSE MRI image to show one month after medical intervention revealed a tiny abscess remnant shaped by debris material. The reference caption matches the model-generated text, yet it uses different word structures. "Sagittal T2-weighted TSE MRI after 1 month of medical treatment: small residual cavity containing some debris at the site of the abscess." These captions show strong similarity, thus proving the model has mastered how to translate essential radiological terms with precise accuracy. The model demonstrates correct interpretation abilities regarding "T2-weighted" and accurately uses these critical diagnostic terms within generated descriptive frameworks, thus validating its ability to process medical imaging features.

A sagittal brain image from MRI shows a select section within the region of interest (ROI) that has been pointed out. A specific marking system guides professional medical practitioners toward areas that might experience persistent effects from the abscess. The model successfully aligns its semantic descriptions with expert human annotator descriptions, thus proving its capability to assist radiologists in reporting and documentation tasks. A detailed assessment spanning numerous imaging scenarios should assess the model's performance behavior, especially when observing anatomical differences or multiple medical conditions and indistinct imaging patterns. These research

findings prove that the model produces clinically significant descriptions for various imaging techniques. Such observational success indicates potential use of this system as an important assisting tool for precise imaging report documentation in medical facilities.

Table 5 : Comparison of proposed model with related work

Method	Datasets	B1	B2	B3	B4	M	R
[17]	IU X-ray	0.50	0.38	0.32	0.28	0.28	0.44
[18]	Mimic CXR	0.68	0.61	0.54	0.48	-	-
[19]	IU X-ray	0.88	0.87	0.87	0.86	-	0.93
[20]	Mimic CXR	0.36	0.24	0.16	0.093	0.32	0.3
[21]	BCD 2018	0.47	0.36	0.27	0.21	0.31	0.46
[22]	ChexPert	0.65	0.50	0.41	0.30	0.42	0.50
[23]	IU X-ray	0.44	0.31	0.22	0.15	-	0.37
[24]	IU X-ray	0.39	0.27	0.19	0.14	0.18	0.33
[25]	Own created	0.56	0.51	0.50	0.49	0.55	0.58
[26]	Stare	0.87	0.66	0.52	0.44	-	-
[27]	Own created	0.27	-	-	0.42	-	-
Proposed Model	**Image set of Radiology**	**0.7959**	**0.7714**	**0.7486**	**0.7300**	**0.6101**	**0.8405**

The proposed model underwent comparison evaluation against established medical image captioning techniques by using BLEU, METEOR and ROUGE metrics, as shown in Table 5. The proposed method achieves better results in various evaluation metrics, but researchers need to acknowledge that the evaluation was conducted through different datasets that have distinct characteristics. The performance metrics on one dataset do not necessarily reveal global superiority of a model because dataset characteristics such as content distribution and annotation standards heavily affect model output quality. Different methods face significant barriers when compared due to dataset structures and annotation styles, which naturally vary between each other. Multiple cited works use IU X-ray as well as MIMIC-CXR together with CheXpert and BCD 2018 datasets that present different levels of image modality and dataset size and textual annotation types. The structured reports of IU X-ray make it hard to compare with the unstructured text descriptions found in MIMIC-CXR. The evaluation scores of the CheXpert and BCD 2018 datasets are affected by annotation variations that occur in length and complexity. The wide range of image types in these tests brings additional difficulties to the comparison process. The data content across different datasets ranges from exclusive chest X-ray images to CT scans, but also includes MRIs together with fundus imaging. Different imaging procedures show different levels of imaging detail, so this distinction plays an essential role because it determines how effective image captioning models become. Exceptional BLEU METEOR ROUGE evaluation results likely stem from the training procedure which worked on a radiology dataset having different diagnostic modalities. The model produces specified performance rates on present testing datasets; however, it fails to maintain parallel results when participating in datasets incorporating various imaging approaches or medical circumstances.

The proposed model retains high semantic and lexical resemblance to reference captions according to B1 = 0.7959, B2 = 0.7714, B3 = 0.7486 and B4 = 0.7300 BLEU scores. The model exhibits successful performance in keeping medical terms intact along with coherent language sequences at all n-gram levels because of its stable BLEU score ratings. The METEOR score of 0.6101 confirms these findings because it evaluates both semantic relationships and necessary clinical terminology detection abilities needed to generate clinical captions. When evaluated through ROUGE scoring, the developed model maintained a

better medical description structure and production sequence, which exceeded reported previous work results to a score of 0.8405.

The implementation of Unified Medical Language System (UMLS) terminology within the dataset resulted in better results for the proposed model. This research utilizes UMLS Concept Unique Identifiers (CUIs) for annotation instead of free-text methods because CUIs establish standardized medical vocabulary in caption outputs. The structured language system enables doctors to use standardized phrases, which decreases the number of ambiguous medical reports. The model's reliability for medical practitioners increases and its interpretability strengthens because the generated captions are linked to official clinical terminologies.

Multiple external datasets including MIMIC-CXR, IU X-ray and CheXpert need to validate the proposed model to check its consistent accuracy and coherence across diverse clinical settings. The complexity of dataset comparison gets worse because of non-standardized structure formats, which indicates the requirement of benchmark datasets that unite annotation practices with unified medical nomenclature and predefined imaging protocols. Future evaluations must include statistical significance analysis by employing confidence intervals with variance metrics to measure the reliability of reported BLEU, METEOR, and ROUGE scores.

Outside of quantitative validation, the true usability of AI-generated captions has to be evaluated in user studies with radiologists to evaluate its effect on real-world radiology workflows, diagnostics efficiency, and workload reduction to assess or approve the deployability of the model into common practice. Solving these issues with cross-dataset benchmarking, statistical validation, and clinical evaluations can lead to AI-driven medical image captioning models that would be robust, interpretable, and ideally integrable into real-world healthcare systems.

Statistical methods were introduced to validate BLEU, METEOR, and ROUGE scores in automated medical image captioning because we wanted to make the result evaluation process stronger. Our 95 % confidence interval calculations verified the stable performance reliability of the model across each metric. The statistical evaluation showed BLEU-1 at 0.7959 ± 0.02, with BLEU-2 at 0.7714 ± 0.02 and BLEU-3 at 0.7486 ±0.03, while BLEU-4 stood at 0.7300 ± 0.03 and METEOR equaled 0.6101 ± 0.025 and ROUGE reached 0.8405 ± 0.02. The BLIP model achieved statistical significance ($p < 0.05$) according to p-value analysis, thus validating its effectiveness for medical image captioning in precise clinical scenarios.

5. Conclusion

The researchers developed an advanced process to create targeted clinical captions through BLIP since they validated the method across various radiology scans including X-rays together with CT and MRI images. BLEU, METEOR and ROUGE scoring methods showed that the model produced captions which matched established expert references by using medical terminology and diagnostic information. The evaluated data demonstrates that artificial intelligence shows promise for radiology practice through its ability to decrease staff workload and enhance operational effectiveness.

This revolutionary system requires several operational obstacles to be solved before it can be implemented at real clinical sites. Primarily, regulatory hurdles and the need for thorough clinical validation pose significant barriers. Medical AI tool deployments require these challenges to meet health standards along with both clinical safety standards and regulatory compliance requirements. The performance of the model would improve by expanding training data to cover multiple imaging methods and healthcare scenarios, which would reduce biased data sets and broaden its clinical implementation capabilities.

Explainable AI methods SHAP and LIME play a vital role in clinician trust building by making the model decision-making process more understandable. The implementation of feature attribution methods along with attention heatmaps provides healthcare staff with visual

tools to see which parts of an image drive the AI's caption generation while helping ensure the confirmation of expert evaluations. Further real-world testing processes must occur to ensure the model delivers executable and dependable medical output whilst overcoming experimental-clinical implementation differences.

Healthcare infrastructures need complete integration of AICT along with regulatory compliance approval and protective data solutions as well as workflow compatibility. Healthcare professionals need to partner with the development team to conduct extensive usability tests which enable feedback to optimize the AI tool according to clinical operational standards and regulatory requirements.

Additional research should concentrate on building the model to perform advanced diagnostic functions that include disease classification identification and abnormality identification with disease progression analysis. This innovation aims to create a sophisticated AI-driven radiology platform which combines automatic diagnostic excellence with hospital system integration, thus enabling mainstream medical imaging adoption through significant operational modifications.

References

[1] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Proceedings of the 39th International Conference on Machine Learning, PMLR 162:12888-12900. DOI: 10.48550/arXiv.2201.12086.

[2] K. Wolff, L. Goldsmith, S. Katz, B. Gilchrest, A. S. Paller, and D. Leffell. Fitzpatrick's Dermatology in General Medicine. McGraw-Hill, New York, NY, USA, 8th edition, 2012.

[3] Zhao, D., Chang, Z., & Guo, S. (2019). A multimodal fusion approach for image captioning. Neurocomputing, 329, 476–485. https://doi.org/10.1016/j.neucom.2018.11.004

[4] Bai, S., & An, S. (2018). A survey on automatic image caption generation. Neurocomputing, 311, 291–304. https://doi.org/10.1016/j.neucom.2018.05.080

[5] Huang, Y., Chen, J., Ma, H., Ma, H., Ouyang, W., & Yu, C. (2022). Attribute assisted teacher-critical training strategies for image captioning. Neurocomputing, 506, 265–276. https://doi.org/10.1016/j.neucom.2022.07.068

[6] Zeng, C., Kwong, S., Zhao, T., & Wang, H. (2022). Contrastive semantic similarity learning for image captioning evaluation. Information Sciences, 609, 913–930. https://doi.org/10.1016/j.ins.2022.07.142

[7] Wang, C., & Gu, X. (2022). Learning joint relationship attention network for image captioning. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2022.118474

[8] Wu, F., Yang, H., Peng, L., Lian, Z., Li, M., Qu, G., Jiang, S., & Han, Y. (2022). Agnet: Automatic generation network for skin imaging reports. Computers in Biology and Medicine, 141. https://doi.org/10.1016/j.compbiomed.2021.105037

[9] Barata, C., Celebi, M. E., & Marques, J. S. (2021). Explainable skin lesion diagnosis using taxonomies. Pattern Recognition, 110. https://doi.org/10.1016/j.patcog.2020.107413

[10] Li, Z. G., & Chen, H. H. Y. C. (2021). Biomedical text similarity evaluation using attention mechanism and siamese neural network. IEEE Access, 9. https://doi.org/10.1109/ACCESS.2021.3099021

[11] Bölücü, N., Can, B., & Artuner, H. (2023). A siamese neural network for learning semantically-informed sentence embeddings. Expert Systems with Applications, 214. https://doi.org/10.1016/j.eswa.2022.119103

[12] Park, H., Kim, K., Park, S., & Choi, J. (2021). Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access, 9, 150560–150568. https://doi.org/10.1109/ACCESS.2021.3124564

[13] Wang, F., Liang, X., Xu, L., & Lin, L. (2022). Unifying relational sentence generation and retrieval for medical image report composition. IEEE Transactions on Cybernetics, 52(6), 5015–5025. https://doi.org/10.1109/TCYB.2020.3026098

[14] Yang, Z., Wang, P., Chu, T., & Yang, J. (2022). Human-centric image captioning. Pattern Recognition, 126. https://doi.org/10.1016/j.patcog.2022.108545

[15] Zhang, Z., Zhang, W., Diao, W., Yan, M., Gao, X., & Sun, X. (2019). VAA: Visual aligning attention model for remote sensing image captioning. IEEE Access, 7, 137355–137364. https://doi.org/10.1109/ACCESS.2019.2942154

[16] Ye, X., Wang, S., Gu, Y., Wang, J., Wang, R., Hou, B., Giunchiglia, F., & Jiao, L. (2022). A joint-training two-stage method for remote sensing image

captioning. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–16. https://doi.org/10.1109/TGRS.2021.3066700

[17] Gajbhiye, G., Nandedkar, A., & Faye, I. et al. (2020). Automatic report generation for chest x-ray images: A multilevel multi-attention approach. In 4th International Conference on Computer Vision and Image Processing, CVIP 2019 (Vol. 1147, pp. 174–182). https://doi.org/10.1007/978-3-030-36152-4_19

[18] Rodin, I., Fedulova, I., & Shelmanov, A. et al. (2019). Multitask and multimodal neural network model for interpretable analysis of x-ray images. In 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019 (pp. 1601–1604). https://doi.org/10.1109/BIBM47256.2019.8983272

[19] Tian, J., Zhong, C., & Shi, Z. et al. (2020). Towards automatic diagnosis from multi-modal medical data. In Interpretability in Machine Intelligence and Medical Image Computing for Multimodal Learning and Decision Support (Vol. 11797, pp. 67–74). https://doi.org/10.1007/978-3-030-50007-5_9

[20] van Sonsbeek, T., Worring, M., & SM, T. et al. (2020). Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest x-rays. In 10th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2020, and the 9th International Workshop on Clinical Image-Based Procedures, CLIP 2020, held in conjunction with the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2020 (Vol. 12445, pp. 106–114). https://doi.org/10.1007/978-3-030-61191-1_12

[21] Yang, S., Niu, J., & Wu, J. et al. (2021). Automatic ultrasound image report generation with adaptive multimodal attention mechanism. Neurocomputing, 427, 40–49. https://doi.org/10.1016/j.neucom.2020.09.084

[22] Yuan, J., Liao, H., & Luo, R. et al. (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, PT VI (Vol. 11769, pp. 721–729). https://doi.org/10.1007/978-3-030-32226-7_80

[23] Yang, S., Niu, J., & Wu, J. et al. (2020). Automatic medical image report generation with multi-view and multimodal attention mechanism. In 20th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2020 (Vol. 12454, pp. 687–699). https://doi.org/10.1007/978-3-030-60248-2_48

[24] Harzig, P., Chen, Y. Y., & Chen, F. et al. (2020). Addressing data bias problems for chest x-ray image report generation. In 30th British Machine Vision Conference, BMVC 2019. https://doi.org/10.48550/arXiv.1908.02123

[25] Syeda-Mahmood, T., Wong, K., & Gur, Y. et al. (2020). Chest x-ray report generation through fine-grained label learning. In 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020 (Vol. 12262, pp. 561–571). https://doi.org/10.1007/978-3-030-59713-9_54

[26] Mishra, S., Banerjee, M., & C., R. et al. (2020). Automatic caption generation of retinal diseases with self-trained RNN merge model. In 7th International Doctoral Symposium on Applied Computation and Security Systems, ACSS 2020 (Vol. 1136, pp. 1–10). https://doi.org/10.1007/978-981-15-2930-6_1

[27] Alsharid, M., El-Bouri, R., & Sharma, H. et al. (2020). A curriculum learning based approach to captioning ultrasound images. In Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis (Vol. 12437). https://doi.org/10.1007/978-3-030-60334-2_8.

[28] Pelka, O., & Friedrich, C. M. (2019). *Radiology Objects in Context (ROCO): A Multimodal Image Dataset*. Zenodo. https://doi.org/10.5281/zenodo.3388203.