Jurnal Masyarakat Informatika, 17(1), 2026
DOI:10.14710/jmasif.17.1.75082
ISSN: 2777-0648

Transformer-Based Encoder-Decoder Model for Medical Image Captioning with Concept Embedding

Husni Fadhilah*, Nugraha Priya Utama
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia
**Corresponding author: 23523034@std.stei.itb.ac.id*

Abstract

This research presents a Transformer-based encoder-decoder model for medical image captioning that incorporates semantic medical knowledge through Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS). The proposed architecture employs a Swin Transformer as the visual encoder and GPT-2 as the language decoder, with CUI integration applied during both caption preprocessing and decoding. Experiments were conducted on the ROCOv2 dataset under two scenarios: baseline (raw captions) and enhanced (CUI-enriched captions). Quantitative evaluation using BLEU, ROUGE, CIDEr, and BERT-based metrics demonstrates that the CUI-integrated model outperforms several baselines, including CNN-LSTM, ViT-BioMedLM, and DeepSeek-VL, achieving a BLEU-1 score of 0.371, ROUGE-L of 0.305, CIDEr of 0.275, and PubMedBERTScore-F1 of 0.893. These results represent a 20.1% improvement in BLEU-1 and a 39.9% increase in ROUGE-L compared to the best-performing model before caption preprocessing (ViT-GPT2 with BLEU-1 = 0.309, ROUGE-L = 0.218). Qualitative assessment by expert radiologists further confirms enhanced diagnostic accuracy, descriptive completeness, and clinical relevance. This study introduces a novel integration of medical semantic knowledge into captioning models, offering a scalable solution for clinical decision support in resource-limited settings such as Indonesia.

Keywords : *Transformer, Medical Image Captioning, Concept Unique Identifier, Unified Medical Language System*

1 Introduction

Radiological imaging has become a cornerstone of modern clinical practice, providing critical insights for diagnosis, treatment planning, and longitudinal disease monitoring. However, the global demand for imaging services has surged at an unprecedented rate, far exceeding the availability of trained radiologists, particularly for modalities requiring complex interpretation, such as CT and MRI. This imbalance has led to heavier clinical workloads, longer turnaround times, and increased susceptibility to diagnostic inaccuracies, often exacerbated by cognitive fatigue and environmental pressures [1].

Recent findings highlight that radiologist burnout, marked by emotional exhaustion and diminished diagnostic performance is closely linked to the growing volume and complexity of imaging studies [1]. To address this, leading institutions such as the Radiological Society of North America (RSNA) have advocated for the integration of artificial intelligence (AI) into radiology workflows. Supported by empirical evidence [2], AI systems offer a promising solution to mitigate these challenges by streamlining image analysis, reducing manual workload, and enhancing diagnostic consistency.

Received: 25 June 2025; Revised: 15 October 2025; Accepted: 21 October 2025; Published: 08 Januari 2026

In particular, deep learning-based approaches have emerged as transformative tools in medical imaging, with the potential to automate interpretation, reduce inter-observer variability, and highlight clinically salient findings [2]. Among these, automated medical image captioning systems aim to generate relevant textual descriptions directly from images, supporting reporting standardization, reducing radiologist workload, and improving clinical decision support across diverse settings [3, 4].

Historically, CNN-LSTM architectures have dominated image captioning tasks. These models utilize convolutional neural networks (CNN) to extract visual features, subsequently employing Long Short-Term Memory (LSTM) networks to sequentially model language generation [5]. While effective for generalized image captioning tasks, CNN-LSTM models present notable limitations, particularly regarding their capability to handle long-range textual dependencies and accurately model complex medical semantics [6]. These models typically struggle with accurately identifying rare pathological findings or generating semantically precise descriptions for uncommon clinical scenarios.

Recent breakthroughs in deep learning have demonstrated that Transformer architectures, first proposed by Vaswani et al. [7], outperform traditional CNN and RNN-based approaches due to their superior ability in capturing global dependencies through self-attention mechanisms. On the vision side, hierarchical designs such as the Swin Transformer [8], extend Vision Transformers (ViT) [9] with shifted-window attention, yielding strong performance on high-resolution radiology images while retaining computational efficiency. For language generation, several decoder-only models are available. GPT-2 [10], GPT-Neo, LLaMA, BART, and T5 among them. GPT-2 remains attractive in research settings because its moderate size (≈ 124 M-774 M parameters) and open licence permit end-to-end fine-tuning. It excels at producing coherent long-form text but also exhibits well-known drawbacks: a 1024-token context limit, a generic pre-training corpus that lacks domain-specific terminology, and a tendency toward factual "hallucination". These weaknesses motivate coupling GPT-2 with external biomedical knowledge bases or switching to heavier encoder-decoder alternatives (BART/T5) that offer stronger factual consistency at the cost of additional compute.

Despite the rapid progress of Transformer-based captioning models, a significant gap remains in the explicit incorporation of structured medical knowledge into these frameworks. The Unified Medical Language System (UMLS) provides a robust foundation to bridge this gap through its Concept Unique Identifiers (CUIs), standardized identifiers linking over 200 biomedical terminologies such as SNOMED CT and ICD-10 [11]. By anchoring generated terms to canonical medical concepts, CUIs reduce synonym drift, resolve lexical ambiguity, and ensure consistency with standardized clinical vocabularies. Prior studies have demonstrated that CUI-based integration enhances both retrieval accuracy [12], named-entity recognition and linking [13], and even zero-shot medical diagnosis [14]. Empirical results further confirm these advantages: the ACapMed model attained a BLEU score of 42.28 % with CUI augmentation compared to 36.28 % using visual features alone [15], while ontology-guided approaches by Zahra and Kate [16], and Zhang et al. [17] achieved higher alignment with domain-standard terminology and improved understanding of clinical intent. This explicit ontology-driven grounding strengthens clinical fidelity and terminological precision, establishing the core novelty and semantic robustness of our proposed Transformer-based captioning framework.

This paper presents a novel Transformer-based encoder-decoder model specifically designed for medical image captioning. The model leverages the Swin Transformer for image encoding and GPT-2 for text decoding, and incorporates pooled Concept Unique Identifier (CUI) embeddings at the vision-language fusion stage. Additionally, it is trained jointly with a caption-cleaning pipeline based

on a large language model (LLM). A key contribution of this approach is the integration of CUIs derived from the Unified Medical Language System (UMLS), which enhances the semantic accuracy and clinical relevance of the generated captions. The proposed model is rigorously evaluated against both a CNN-LSTM baseline and the state-of-the-art ViT-BioMedLM model [18], demonstrating improved performance across multiple metrics and underscoring its potential clinical applicability.

2 Literature Review

Medical image captioning has evolved through several architectural paradigms, each contributing to incremental improvements in semantic alignment and textual fluency. Early works predominantly relied on CNN-RNN architectures, where convolutional neural networks (CNNs) extract spatial features from images, and recurrent neural networks (RNNs) or long short-term memory (LSTM) units generate captions sequentially. Notable examples include the Show-Attend-Tell model [19], which introduced soft visual attention mechanisms, and the ACapMed system [15], which tailored CNN-LSTM pipelines to the biomedical domain. While effective in generating syntactically correct captions, these models struggle to capture complex semantic dependencies and often underperform in handling rare or clinically nuanced terminologies, limiting their clinical interpretability.

To improve terminological accuracy and reduce hallucinations, template-based and phrasebook models were proposed, such as PhraseAug [20], which augment captions using structured medical vocabularies. Although these models increase precision and domain conformity, they rely heavily on pre-defined structures, sacrificing flexibility and reducing the ability to generalize to previously unseen or ambiguous clinical conditions. This rigidity hinders their scalability in real-world applications involving diverse pathologies.

With the advent of Transformer-based architectures, a significant shift occurred in both visual and textual representation learning. Vision Transformers (ViT) [9] and Swin Transformers [8] replaced convolutional hierarchies with self-attention, enabling global receptive fields and hierarchical spatial encoding, which are particularly beneficial for parsing high-resolution radiology images. Simultaneously, autoregressive language models such as GPT-2 [10] and encoder-decoder models like BART and T5 have demonstrated the ability to generate coherent and contextually rich narratives, thanks to their multi-layer self-attention mechanisms that preserve long-range dependencies. These models have been applied to radiology report generation, showing improved narrative coherence, clinical phrasing, and alignment with expert-written summaries [21].

Beyond domain-specific transformer applications, recent advances in large-scale vision–language pre-training have produced powerful multimodal models capable of unified understanding and generation. BLIP [37] introduced a unified encoder–decoder framework that bootstraps noisy image–text pairs for improved captioning and retrieval. DeepSeek-VL [38] extended this paradigm to real-world multimodal reasoning using a hybrid high-resolution vision encoder and balanced cross-modal training strategy. Qwen2-VL [36] further enhanced perceptual flexibility through dynamic resolution processing and multimodal rotary position embeddings, enabling robust comprehension across diverse image and video tasks. These developments illustrate the transition toward high-resolution, general-purpose vision–language architectures that inspire domain-specific adaptations such as the proposed Swin Transformer–GPT2 framework for radiology captioning.

Despite these advances, most Transformer-based captioning systems remain data-driven, relying on surface-level co-occurrence patterns rather than grounding output in medical ontologies. This leads to semantic drift and poor handling of synonyms, abbreviations, or context-specific terminology. To address this, recent efforts have focused on domain-specific vision-language models (VLMs), such as BioMedLM [22], SAPBERT [23], and UmlsBERT [24], which enrich biomedical embeddings using external knowledge bases like the Unified Medical Language System (UMLS). For example, UmlsBERT introduces semantic type embeddings to group medically related terms, while SAPBERT performs self-alignment on synonymous biomedical entities across vocabularies, enhancing NER and relation extraction tasks.

However, while these models excel in classification, retrieval, and entity linking, their application to generative tasks such as captioning remains limited. Notably, research by Zhang et al. [25] demonstrated that integrating UMLS-based graphs can significantly improve zero-shot medical diagnosis. Similarly, Beam et al. [26] introduced Cui2Vec, an embedding trained on large-scale multimodal data, showing performance gains in clinical prediction tasks. Yet, few studies have systematically integrated Concept Unique Identifiers (CUIs) into end-to-end vision-to-text captioning pipelines. Therefore, the integration of CUI representations within Transformer-based captioning architectures presents a novel and promising research direction. By anchoring generated captions to standardized medical concepts, such an approach offers the potential to improve not only lexical fluency but also clinical correctness, terminological consistency, and semantic interpretability key factors for real-world deployment in healthcare environments.

2.1 Evaluation Metrics

This study used five widely accepted evaluation metrics to assess the syntactic (BLEU, ROUGE, CIDEr) and semantic (BERTScore, PubMedBERTScore) quality of generated captions:

1) BLEU

BLEU is a widely used automatic evaluation metric that measures the n-gram precision between a generated caption and one or more reference captions, while incorporating a brevity penalty (BP) to discourage overly short outputs that might achieve high precision by omitting content. By evaluating overlapping n-grams of varying lengths (typically from unigrams to four-grams), BLEU captures how closely the generated text aligns with reference expressions at the lexical level. Measures n-gram precision with a brevity penalty [31]. The BLEU score is computed as defined in equation (11).

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log\left(p_n\right)\right) \quad (11)$$

where BP is the brevity penalty, p_n is the modified n-gram precision, and w_n are weights typically set to uniform distribution.

2) ROUGE-L

ROUGE-L evaluates caption quality by measuring the Longest Common Subsequence (LCS) between the generated caption and the reference caption. Unlike n-gram-based metrics, ROUGE-L captures sentence-level structural similarity without requiring the matched words to be contiguous, making it more robust to variations in phrasing while preserving word order. Evaluates the longest sequence of words that appear in both captions in the same order [32], though not necessarily

contiguously. The recall, precision, and F-measure components of ROUGE-L are formally defined in equations (12) – (14).

$$Recall\ (R_{LCS}) = \frac{LCS(c,\ g)}{m} \tag{12}$$

$$Precision\ (P_{LCS}) = \frac{LCS(c,\ g)}{n} \tag{13}$$

$$ROUGE - L\ (F_{LCS}) = \frac{(1+\beta^2)\ x\ R_{LCS}\ x\ P_{LCS}}{R_{LCS} + P_{LCS}\ x\ \beta^2} \tag{14}$$

where LCS is longest common subsequence of candidate caption (c) with length m and ground-truth caption (g) with length n. To ensuring a balanced trade-off between precision and recall, $\beta = \frac{P_{LCS}}{R_{LCS}}$ when $\frac{F_{LCS}}{R_{LCS}} = \frac{F_{LCS}}{P_{LCS}}$.

3) CIDEr

CIDEr is designed to measure the consensus between a generated caption and a set of reference captions by employing TF–IDF-weighted n-gram representations. Unlike BLEU and ROUGE, CIDEr emphasizes n-grams that are frequent within a specific caption but relatively rare across the entire dataset. Measures consensus between the generated caption and a set of reference captions using TF-IDF weighting of n-grams, which reflects content relevance and diversity [33]. The CIDEr score is computed according to equation (15)

$$CIDEr\ Score = \sum_{n=1}^{N} w_n \left(\frac{1}{m} \sum_j \frac{g^n(c_i)\ x\ g^n(g_{ij})}{\left|\left| g^n(c_j) \right|\right| x \left|\left| g^n(g_{ij}) \right|\right|} \right) \tag{15}$$

where the formula represents the TF-IDF vector for the n-grams in the candidate text $(g^n(c_i))$ and reference text $(g^n(g_{ij}))$.

4) BERTScore

BERTScore evaluates caption quality at the semantic level by leveraging contextualized token embeddings obtained from large pre-trained transformer models, such as microsoft/deberta-xlarge-mnli. Instead of relying on exact word overlap, BERTScore computes pairwise cosine similarities between tokens in the generated caption and the reference caption, enabling the capture of paraphrasing and semantic equivalence [34]. The cosine similarity between token embeddings is defined in equation (16), while token-level precision, recall, and F1-score are computed using equations (17) – (19) :

$$Cosine\ Similarity(\hat{x}_j, x_i) = \frac{\hat{x}_j^\top x_i}{||\hat{x}_j|| \cdot ||x_i||} \tag{16}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} max_{x_i \in x} \hat{x}_j^\top x_i \tag{17}$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{18}$$

$$F_{BERT} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{19}$$

where x denote the reference text and $\hat{x}$ the predicted text. The terms $||\hat{x}_j||$ and $||x_i||$ represent the norms (or lengths) of their respective embedding vectors. Here, x_i and $\hat{x}_j$ refer to the embedding vectors of the i-th and j-th tokens from the reference and predicted texts, respectively, and the similarity between them is computed using cosine similarity.

5) PubMedBERTScore (F1)

PubMedBERTScore extends BERTScore by employing domain-specific biomedical embeddings derived from the PubMedBERT model, such as microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract. While maintaining the same precision, recall, and F1-score formulation defined in Equations (17)–(19), PubMedBERTScore operates within a biomedical embedding space that more accurately represents clinical terminology and domain-specific semantics [35], which better captures clinical semantics.

3 Research Methods

This section describes the proposed methodology for medical image captioning, which integrates visual feature extraction and language generation within a unified encoder–decoder framework. The model is designed to effectively capture both hierarchical visual representations from radiology images and domain-specific semantic information required for generating clinically meaningful captions. The overall architecture of the proposed model is illustrated in Figure 1, which depicts the integration of a Swin Transformer as the image encoder and GPT-2 as the autoregressive text decoder, augmented with Concept Unique Identifier (CUI) embeddings derived from UMLS.

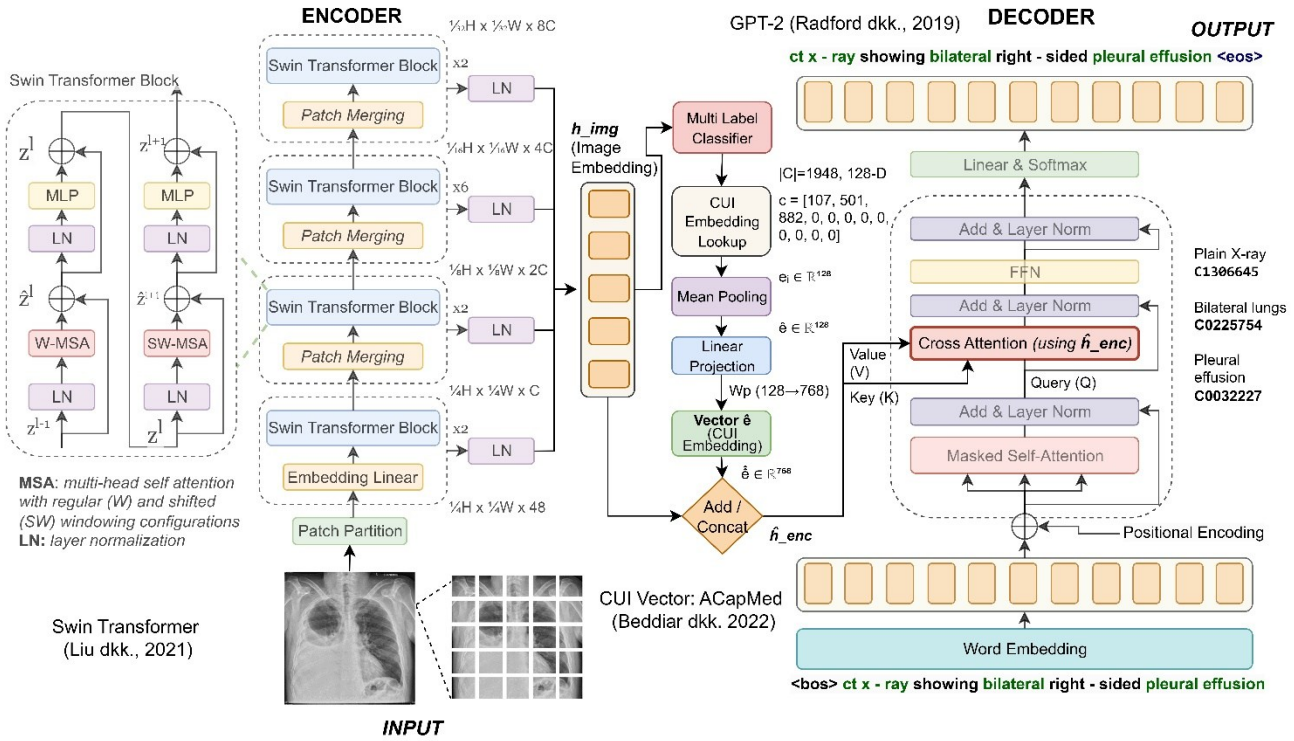


Figure 1 The proposed architecture for medical image captioning integrates Swin Transformer as the image encoder and GPT-2 as the autoregressive decoder, augmented with Concept Unique Identifier (CUI) embeddings from UMLS to enhance semantic grounding. The encoder extracts hierarchical features from radiology images, while the decoder generates text using both image embeddings and CUIs via cross-attention.

This study employs a Transformer-based model with encoder-decoder architecture tailored for medical image captioning. The encoder utilizes the Swin Transformer, which effectively captures hierarchical visual features by dividing the input image into non-overlapping patches and progressively merging them through successive transformer layers. This hierarchical representation allows the model to extract multi-scale visual context critical for understanding complex medical imagery.

The extracted visual features are then passed to a GPT-2 decoder, an autoregressive language model capable of generating coherent and semantically rich captions. The encoder and decoder are integrated using HuggingFace's *VisionEncoderDecoderModel*, where cross-attention mechanisms enable the decoder to attend to relevant visual information during caption generation. This setup ensures effective fusion of visual and textual modalities for producing accurate and context-aware medical descriptions.

3.1 Dataset and Preprocessing

The performance of medical image captioning models is strongly influenced by the quality, diversity, and semantic alignment of the underlying data, particularly in clinical domains where visual patterns must be accurately mapped to meaningful textual descriptions. The inclusion of standardized Concept Unique Identifier (CUI) annotations enables explicit semantic grounding of captions in biomedical knowledge, which is essential for improving clinical relevance and interpretability. An overview of representative samples from the dataset, illustrating medical images paired with their corresponding captions and annotated CUIs, is presented in Figure 2 to provide qualitative insight into the diversity and structure of the data used in this work.

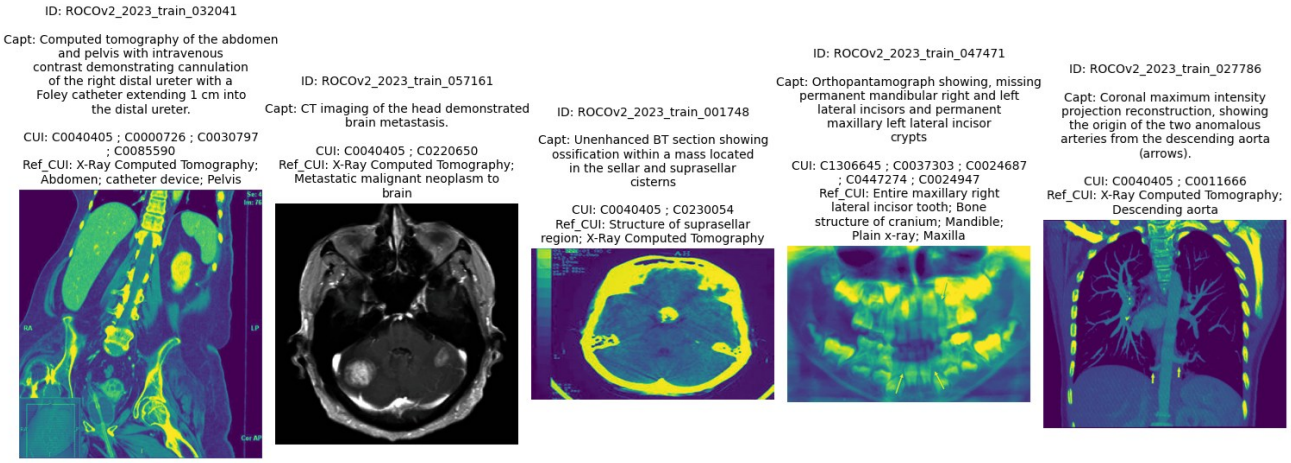


Figure 2 Sample images and captions from the ROCOv2 dataset [18] with annotated Concept Unique Identifiers (CUIs). The examples span multiple modalities including CT scan, MRI, X-ray, and ultrasonography, demonstrating the diversity of anatomical regions and clinical contexts. Each caption is paired with its corresponding CUIs and reference concepts.

This study utilized the ROCOv2 (Radiology objects in context version 2) [18] dataset comprising 80.080 samples (60.163 train, 9.945 validation, and 9.972 test) across X-ray, CT, MRI, PET, Ultrasound, Angiography, Mammography, and Fluoroscopy modalities. Images were resized, normalized, and segmented into patches compatible with Swin Transformer inputs. MedCAT facilitated the annotation of CUIs in captions, aligning concepts to token spans. Dataset filtering was applied to focus on chest and brain anatomical regions, ensuring consistency and relevance in evaluations.

3.2 Image Preprocessing

Prior to model training, all medical images were processed through a standardized preprocessing pipeline to ensure input consistency and improve training stability across diverse imaging modalities. Images were first resized to a uniform spatial resolution to meet the fixed input requirements of the vision encoder, thereby minimizing variability introduced by heterogeneous acquisition settings. Intensity normalization was subsequently applied to harmonize pixel value distributions, facilitating stable gradient propagation and more efficient model convergence. Following these steps, images were segmented into non-overlapping patches compatible with the Swin Transformer architecture, enabling hierarchical feature extraction through localized self-attention and progressive patch merging. This preprocessing strategy ensures seamless integration with the Swin Transformer while preserving clinically relevant visual information essential for accurate medical image captioning.

3.3 Caption and CUI Annotation

To enhance semantic grounding and domain-specific understanding, this study incorporate Concept Unique Identifiers (CUIs) extracted from the captions using MedCAT [27], which annotates medical concepts based on the Unified Medical Language System (UMLS). Tokenization is applied to align CUIs with the corresponding textual spans in captions. The CUIs are incorporated using an early fusion strategy, in which they are appended directly to the caption text during training. This method reinforces the model's understanding of medical terminology and contributes to improved clinical relevance and interpretability of the outputs.

3.4 Caption Cleaning with LLM-based Pre-processing

To ensure that textual inputs are syntactically clean, clinically precise, and semantic consistency, all raw captions were normalised with a large-language-model (LLM) pipeline based on *google/gemini-2.5-flash-preview-05-20*. The detailed prompt is provided in the Appendix.

1) Cleaning Stage Objectives

The cleaning stage pursues three objectives:

a. Noise removal: strip non-ASCII artifacts, figure references, case IDs, and citations.
b. Clinical focus: retain anatomical/pathological findings and quantitative measures while discarding procedural metadata.
c. Semantic alignment: leverage UMLS CUIs as context so that medically relevant terms are preserved or clarified.

2) API workflow

For each (image, caption, CUI) triple this study issue a single LLM call with a structured prompt. The LLM returns a "cleaned caption" that is (i) one-two sentences, (ii) free of irrelevant tokens, and (iii) still anchored to the supplied CUIs. This study cache the response and fall back to the original caption if the LLM score fails a simple length/ASCII sanity check.

The automated caption cleaning step substantially reduces the presence of outliers and overly long captions across all dataset splits. As illustrated in the histograms in Figure 3, the average caption length in the merged set decreases from approximately 22 words to 18 words after cleaning. This corresponds to a reduction of approximately 26% in length variance, effectively producing

more uniformly distributed caption lengths while preserving essential clinical semantics. The sharper distribution peak and lower tail frequency indicate that the processed captions are not only shorter on average but also more consistent across the training, validation, and test sets. The construction of this prompt, including its constraints and formatting rules, is detailed in Algorithm 1

Algorithm 1: LLM-Based Prompt Construction for Medical Caption Cleaning

Input: Original Caption $\mathcal{C}$, Concept Unique Identifiers (CUIs) $\mathcal{U}$
Output: Cleaned Caption $\mathcal{C}^*$
Function BuildPrompt($\mathcal{C}$, $\mathcal{U}$):
1. Initialize prompt with: "You are a medical language processing assistant..."
2. Append the following instructions:
  a) Fix corrupted or non-standard characters (e.g., â–ª, Ã—, Â)
  b) Remove references to figures, years, case numbers, visual annotations, citations
  c) Remove procedural statements unrelated to image appearance
  d) Keep anatomical/pathological findings and quantitative values
  e) Rewrite into 1-2 concise, medically accurate sentences
  f) Enrich vague captions with common visual context (no hallucination)
  g) Normalize spacing and punctuation; expand ambiguous abbreviations
  h) Use provided CUIs to clarify terminology
3. Append: Original Caption: "{$\mathcal{C}$}"
4. Append: CUIs: "{$\mathcal{U}$}"
5. Append: Cleaned Caption:
Return the constructed prompt $\mathcal{P}$

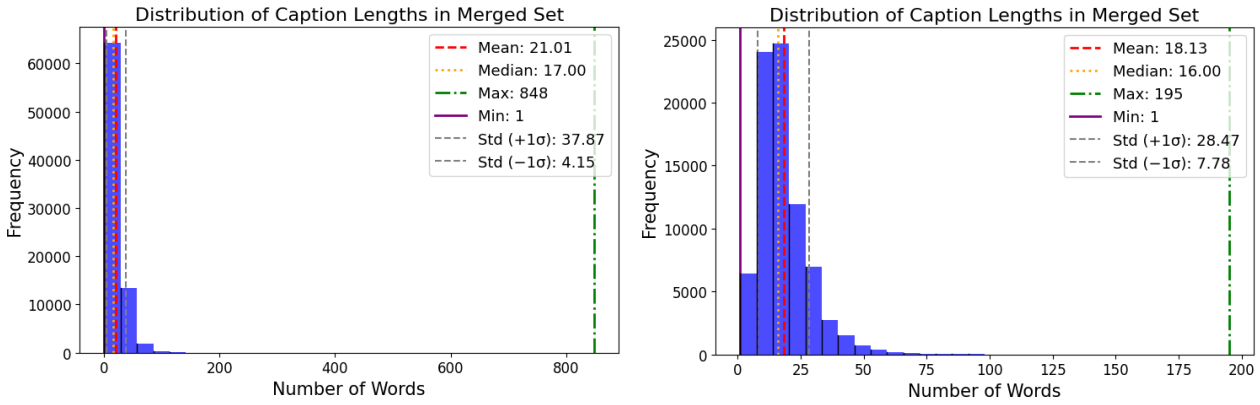


Figure 3. Distribution of caption lengths across merged sets (training, validation, test) before (left) and after (right) automated cleaning. The red dashed line indicates the mean length.

3) Justification for GPT-2 Decoder and Role of Gemini

This study adopts GPT-2 as the generative decoder to ensure efficient, controllable, and reproducible autoregressive caption generation within the proposed encoder–decoder framework. GPT-2 provides open and fine-tunable model weights, enabling seamless integration with the visual encoder while avoiding the computational overhead associated with full encoder–decoder architectures such as BART or T5. In contrast, Gemini 2.5 Flash is intentionally restricted to the preprocessing stage for caption cleaning and is not employed as a generative component during model training. Although Gemini exhibits strong instruction-following capabilities for text refinement, its non-deterministic reasoning behavior, limited transparency, and lack of support for local fine-tuning make it less suitable for use as a trainable decoder in this context [28]. Instead, its role is confined to offline caption normalization to improve dataset consistency without introducing variability into the learning process. A compact comparison of Gemini 2.5 Flash with other large language models, including GPT-4o and Claude 3.5 Sonnet, is summarized in Table 1, highlighting

their relative suitability for caption-cleaning tasks in terms of controllability, deployment constraints, and computational efficiency [29].

Table 1 A Comparison of Large Language Models: Gemini 2.5 Flash, GPT-4o, and Claude 3.5 Sonnet

Model	Reasoning / Multimodal Strength	Context Window	Speed / Latency	Notes / Cost Considerations
Gemini 2.5 Flash	Strong reasoning & multimodal	~1,048,576 tokens	~274 tokens/s, moderate latency	Input ≈ $0,10/M token, Output ≈ $0,40/M token. Lower cost per token, suited for batch cleaning
GPT-4 / 4.1	Very strong reasoning	~128K – 1M tokens	Higher latency (~0.45 s)	Input ≈ $2,5/M token, Output ≈ $10/M token. High cost, slower for large batch processing
Claude 3.5 Sonnet	Excellent reasoning and safety	~64–200K tokens	Efficient latency	Input ≈ $3/M token, Output ≈ $15/M token. Highest cost, but less context capacity

3.5 Concept Unique Identifier (CUI) Integration

While the Swin Transformer-GPT2 backbone captures visual and linguistic context, radiology captions also require precise medical terminology. Inspired by recent evidence that embedding UMLS concepts can improve semantic accuracy in clinical NLP [20, 21, 22] this study incorporate CUIs into both the training and generation stages. Figure 2 highlights the additional CUI pathway in orange.

1) CUI Extraction and Vocabulary Construction

From the *Ref_CUI* column in ROCOv2 this study parse every CUI string, split on the pipe delimiter, deduplicate, and build a lookup table CUI → index. The final vocabulary contains $|C| = 1948$ unique CUIs (max-frequency cutoff < 5 removed). Each CUI receives an embedding $\mathrm{e_c} \in \mathrm{R}^{128}$ initialised uniformly at random.

2) Dataset Encoding

For every image-caption pair, this study create a fixed-length tensor $\mathrm{c} \in \mathrm{N}^{\mathrm{MAX\_CUI}}$ that stores the CUI indices occurring in the caption; unused slots are padded with 0. The value of $\mathrm{MAX\_CUI} = 12$ (95$^{\mathrm{th}}$ percentile of ROCOv2).

3) Model Augmentation

The vanilla *VisionEncoderDecoderModel* is extended with:

a. A Multi-label Classifier that takes the pooled output of the vision encoder to predict relevant CUIs using a linear layer and a ReLU activation function. This classifier allows the model to predict CUIs from the image itself during inference.
b. CUI Embedding Layer: that maps c to a pooled vector $\hat{e} = \mathrm{mean}[\mathrm{e_{c_i}}]$
c. Projection Layer $W_p \in R^{128\,x\,d_{enc}}$ with the encoder hidden size $d_{enc} = 768$
d. Fusion Mechanism: that merges the projected CUI vector $\hat{e} = \hat{e}W_p$ with the CLS-token of the Swin encoder, either by simple addition *(add)* or channel-wise concatenation *(concat)*.

4) Forward Pass

Given pixels x and CUI indices c:

$$h_{img} = E_{Swin}(x) \quad (1)$$

$$\hat{e} = f_{proj}(Embed(c)) \quad (2)$$

$$\hat{h}_{enc} = Fuse(h_{img}, \hat{e}) \quad (3)$$

$$y = D_{GPT2}(\hat{h}_{enc}, y < t) \quad (4)$$

where D_{GPT2} is the autoregressive decoder with cross-attention to $\hat{h}_{enc}$.

5) Training and Inference

Training is performed using standard cross-entropy loss on caption tokens, with c provided as an additional input. During inference, the generate() function is overridden to inject the tensor c corresponding to the test image, ensuring CUI guidance during decoding. Upon completion of training, the Swin parameters, GPT-2 weights, CUI embeddings, tokenizer, and feature extractor are saved as a unified artefact.

The proposed pipeline grounds language generation in a structured ontology, aiming to minimise synonym drift and improve clinical fidelity of the produced reports. Ablation results in Table 7 confirm that CUI fusion yields consistent gains across BLEU-1, ROUGE-L, CIDEr, and BERTScore.

6) Examples of Cleaned Captions

Table 2 Examples of Raw and Cleaned Captions, Demonstrating the Effectiveness of LLM-Based Cleaning Pipeline in Retaining Clinical Focus

ID	Original Caption	Cleaned Caption	Ref CUI
ROCOv2 train 000389	Radiograph of an artificially decalcified rib, with 54.7% of the calcium removed. From: Lachman E and Whelan M.A: The roentgen diagnosis of osteoporosis and its limitations. Radiology 26, 165–177 (1936) (with permission).	Chest x-ray of an artificially decalcified rib, with 54.7% of the calcium removed, demonstrating features of osteoporosis.	Plain x-ray; Chest; Osteoporosis
ROCOv2 train 001380	Mediorenal tumoral mass classified as T1, suggestive for RCC (‘Fundeni’ Archives)	Computed Tomography demonstrating mediorenal tumoral mass classified as T1, suggestive for renal cell carcinoma.	X-Ray Computed Tomography
ROCOv2 train 002427	Initial chest x-ray on presentation to the emergency department. Chest x-ray showcasing patchy ground-glass opacifications	Initial chest x-ray on presentation to the emergency department showcasing patchy ground-glass opacifications.	Anterior-Posterior; Plain x-ray; Chest
ROCOv2 valid 005053	Myometrial thickness (red line) in early pregnancy was 7 mm in case no. 5 from group B, which had a normal placenta during late pregnancy.	Myometrial thickness in early pregnancy was 7 mm, with a normal placenta during late pregnancy.	Ultrasonography; Pregnancy
ROCOv2 valid 005598	Another case of a known arteriovenous fistula status post Onyx embolization	Angiogram demonstrating an arteriovenous fistula status post Onyx embolization.	Arteriovenous fistula; angiogram
ROCOv2 test 000017	HRCT done on presentation to the ER that shows B/L ground glass infiltrates with patchy consolidations involving mainly the peripheries	High-resolution computed tomography shows bilateral ground-glass infiltrates with patchy consolidations, predominantly involving the peripheries.	X-Ray Computed Tomography
ROCOv2 test 000158	CT scan of the chest.CT scan of the chest showing scattered reticular, ground-glass, atelectatic and fibrotic changes again seen in both lungs. These are slightly worsened compared to Figure 1 especially in the right upper lobe where there is a groundglass patchy infiltrate of 5 cm in size with associated new cavity of 2 cm in the right middle lobe (blue arrow). .	Computed tomography of the chest shows scattered reticular, ground-glass, atelectatic, and fibrotic changes in both lungs, with slight worsening in the right upper lobe, where there is a 5 cm ground-glass patchy infiltrate with an associated 2 cm new cavity in the right middle lobe.	Atelectatic; Cavitation; Structure of middle lobe of right lung; X-Ray Computed Tomography; Structure of right upper lobe of lung; Bilateral lungs

Table 2 illustrates examples of raw image captions from the ROCOv2 dataset and their corresponding cleaned versions generated via the LLM-based prompt cleaning procedure (see Algorithm on Table 1. The cleaning process removes extraneous references (e.g., case numbers,

citations) and reinforces medically relevant content, particularly by aligning terminology with the reference CUIs.

7) Dataset Filtering

To maintain consistency and clinical relevance, the dataset was filtered based on anatomical regions, primarily focusing on chest and brain imaging studies. This strategic selection allowed the evaluation to concentrate on common diagnostic scenarios, thereby ensuring that the generated captions met practical clinical standards and could be effectively evaluated for semantic and diagnostic accuracy.

8) Mathematical Formulation

The image captioning task is modeled as a conditional sequence generation problem. Given an image I, the model aims to generate a caption $Y = \{y_1, y_2, \dots, y_T\}$ by maximizing the conditional likelihood:

$$\hat{Y} = \arg max_{Y,} P(Y \mid I; \theta) \quad (5)$$

where θ denotes the parameters of the encoder-decoder network.

The encoder E maps the input image I into a sequence of visual feature representations V:

$$V = E(I) \quad (6)$$

In the implementation, E is a Swin Transformer that extracts hierarchical patch embeddings and models local-global visual dependencies via window-based self-attention.

The decoder D, based on GPT-2, generates each word y_t conditioned on previously generated tokens and the visual features V:

$$P(y_t \mid y_{<t}, V) = D(y_{<t}, V) \quad (7)$$

The attention mechanism used in both encoder and decoder layers is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (8)$$

where Q, K, and V are the query, key, and value matrices derived from input embeddings, and d_k is the dimensionality of the key vectors.

In the decoder's cross-attention, Q is derived from the decoder's token embeddings, while K and V are computed from the encoder's visual outputs. This mechanism allows the model to align image context with generated words.

The model is trained using the cross-entropy loss function over the caption sequence:

$$L(\theta) = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, V; \theta) \quad (9)$$

When Concept Unique Identifiers (CUIs) are integrated, this study introduce an additional embedding vector $\bar{c}$, representing pooled semantic information from UMLS concepts. This embedding is concatenated with or added to the encoder features before decoding:

$$P(y_t \mid y_{<t}, V; \bar{c}) = Decoder(y_{<t}, V; \bar{c}) \quad (10)$$

This formulation allows the model to leverage both visual and structured semantic information, improving the clinical accuracy and terminological precision of generated captions.

3.6 Hyperparameters and Training Details

The proposed model was implemented using the HuggingFace Transformers library by integrating a Swin Transformer encoder and a GPT-2 decoder through the VisionEncoderDecoderModel interface. Model training was carried out using PyTorch with mixed-precision (FP16) enabled to improve memory efficiency and computational throughput. The key training and hyperparameter tuning configurations adopted in this study, including optimizer settings, learning rates, batch sizes, and training schedules, are summarized in Table 3.

Table 3 Summary of Training and Tuning Configurations

Parameter	Setting	Parameter	Setting
Framework	PyTorch + HuggingFace	**Fusion Mechanism**	Additive (CUI + visual embeddings)
Encoder	Swin Transformer (Base, IN-22k)	**Decoder**	GPT-2 (12 layers, 768 hidden, 12 heads)
CUI Embedding Dim.	128 → projected to 1024	**Input / Caption Length**	224×224 px / 128 tokens
Max CUIs per Sample	12	**Batch Size / Epochs**	8 (accum.) / 10
Optimizer / LR	AdamW (β_1=0.9, β_2=0.999), 5 × 10^{-5} LR	**LR Schedule**	Linear decay + 10 % warm-up
Regularization	Dropout 0.1, Grad clip 1.0	**Early Stopping Metric**	ROUGE-L (patience = 3)
Model Selection	Highest validation BERTScore-F1	**Hardware**	NVIDIA RTX A5000 (24 GB), 45–60 min / epoch

a. Encoder and Decoder

The Swin Transformer encoder adopts a hierarchical window-based attention mechanism. This study employed the *microsoft/swin-base-patch4-window7-224-in22k* variant pretrained on ImageNet-22k, comprising 24 Transformer blocks (2 + 2 + 18 + 2) across four stages. Hidden dimensions increase progressively from 128 to 1024, producing a final feature map of size 1/32 H × 1/32 W × 1024. The GPT-2 decoder (*openai-community/gpt2*) consists of 12 Transformer blocks with 768 hidden units and 12 attention heads. A linear projection layer aligns the encoder and decoder dimensions. Both components were frozen during initial warm-up epochs, then jointly fine-tuned.

b. CUI Embedding

Each Concept Unique Identifier (CUI) is mapped to a learnable 128-dimensional embedding, pooled and projected to the encoder hidden size before additive fusion. This strategy yielded the highest validation BERTScore-F1 among tested configurations.

c. Optimization and Learning Schedule

Training used the AdamW optimizer ($\beta_1 = 0.9, \beta_1 = 0.999$, weight decay = 0.01) for 10 epochs, with a learning rate of $5x10^{-5}$, linear decay, and 10 % warm-up. Gradient clipping (threshold = 1.0) stabilized updates.

d. Early Stopping and Model Selection

Training used early stopping based on ROUGE-L (patience = 3 epochs). The final checkpoint was selected using the highest validation BERTScore-F1.

e. Other Configurations

All input images were resized to 224 × 224 px, and captions were tokenized with a maximum length of 128 tokens. Each sample included up to 12 CUIs (zero-padded if fewer). The effective batch size was 8 with gradient accumulation, and a dropout rate of 0.1 was used in attention and feedforward layers. Experiments were conducted on a NVIDIA RTX A5000 (24 GB VRAM), with an average training time of 45–60 minutes per epoch depending on dataset size and augmentation strategy. Label smoothing (ε = 0.1) was tested but omitted due to negligible gains.

4 Results and Discussion

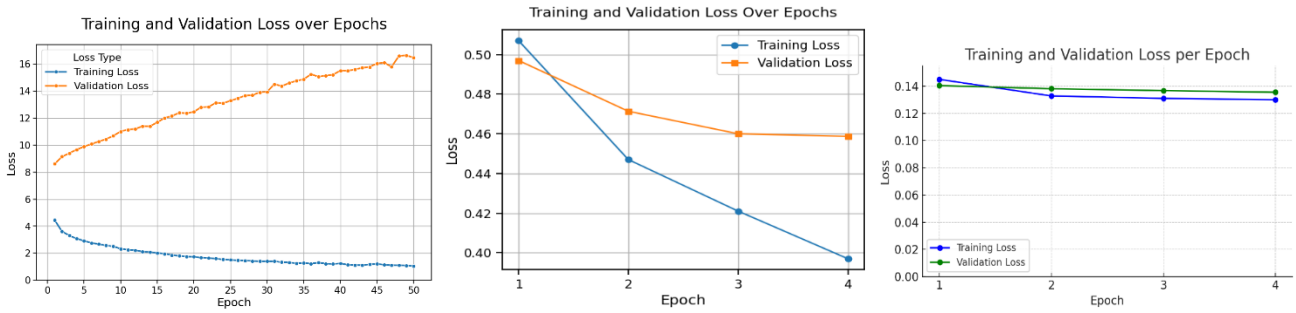
4.1 Evaluation Setup and Metrics

To evaluate the proposed Swin Transformer-GPT2 model with CUI integration, experiments were conducted using the ROCOv2 dataset, comparing against two baselines: (1) a CNN-LSTM model using EfficientNet as the visual encoder, and (2) a pretrained ViT-BioMedLM vision-language model. All models were trained and tested using the same dataset split and caption preprocessing pipeline for fairness.

4.2 Quantitative Results

1) Main Results

To analyze the training dynamics and convergence behavior of the compared models, the evolution of training and validation loss across epochs is examined. A comparative visualization of the loss curves for EfficientNetB0–LSTM, Swin Transformer–GPT-2, and DeepSeekVL 1.3B Chat is presented in Figure 4, providing insight into model stability and learning efficiency throughout the training process.



(a) (b) (c)

Figure 4 Training and validation loss over epoch. From left to right: (a) EfficientNetB0-LSTM, (b) Swin Transformer-GPT2, (c) DeepSeekVL 1.3b Chat.

Tables 5 and 6 present the evaluation results across all metrics for the compared models. The Swin Transformer–GPT-2 + CUI model consistently achieves the highest performance across all evaluation metrics, significantly outperforming the baseline approaches. Table 4 summarizes the experimental results obtained before caption preprocessing, with the best score for each metric

highlighted in bold, while Table 5 reports the corresponding results after the caption cleaning stage, enabling a direct comparison of the impact of preprocessing on model performance.

Table 4 Summary of Experimental Results (Before Caption Preprocessing). The best score for each metric is highlighted in bold.

Model	BLEU-1	ROUGE-L	CIDEr	BERTScore (F1)	PubMedBERTScore (F1)
EfficientNetB0-LSTM	0.228	0.146	0.052	0.628	0.876
EfficientNetB7-LSTM	0.229	0.161	0.057	0.636	0.877
Swin Transformer Base-GPT2	**0.315**	**0.227**	0.139	**0.669**	**0.890**
ViT Base-GPT2	0.309	0.218	**0.152**	0.664	0.889
DeepSeek-VL 1.3B Chat	0.276	0.211	0.103	0.660	0.886
ViT-BioMedLM	0.276	0.185	0.063	0.645	0.881

Table 5 Summary of Experimental Results (After Caption Preprocessing). The best score for each metric is highlighted in bold.

Model	BLEU-1	ROUGE-L	CIDEr	BERTScore (F1)	PubMedBERTScore (F1)
Swin Transformer Base-GPT2 + CUI	**0.371**	**0.305**	**0.275**	**0.719**	**0.893**
ViT Base-GPT2 + CUI	0.364	0.298	0.288	0.716	0.892
Qwen2VL 2B	0.343	0.269	0.166	0.693	0.882
BLIP-Base	0.297	0.268	0.135	0.695	0.887
DeepSeek-VL 1.3B Chat	0.295	0.273	0.184	0.698	0.891
ViT-BioMedLM	0.285	0.209	0.124	0.689	0.882

2) Performance Analysis

A comparative evaluation of model performance before and after caption preprocessing reveals substantial improvements across all evaluation metrics, particularly when Concept Unique Identifier (CUI) embeddings from UMLS are incorporated. The experiments included six models in the “before cleaning” phase and six corresponding or enhanced variants in the “after cleaning” phase.

a. Before Preprocessing: Transformer Models Lead, but With Limits

In the initial phase (before caption cleaning), Swin Transformer Base-GPT2 emerged as the most semantically coherent model, achieving a BERTScore (F1) of 0.669 and a CIDEr score of 0.139. Similarly, ViT Base-GPT2 recorded a CIDEr score of 0.152, slightly higher than Swin-GPT2, but with lower BLEU-1 (0.309 vs. 0.315). These results suggest that both models are capable of capturing clinical semantics and generating coherent sentences, but still suffer from noise and inconsistency in unprocessed training data.

Traditional CNN-LSTM baselines, such as EfficientNetB0-LSTM and EfficientNetB7-LSTM, showed the lowest scores across all metrics. Their CIDEr scores (0.052 and 0.057) and BERTScore F1 (0.628-0.636) indicate limited capacity in representing complex medical concepts or aligning well with reference descriptions. This reinforces the limitations of recurrent models in medical image captioning, especially when dealing with domain-specific language variability and sparse findings. Interestingly, DeepSeek-VL 1.3B Chat, a large-scale vision-

language model pretrained on diverse multimodal data, demonstrated reasonable scores in both BLEU-1 (0.276) and ROUGE-L (0.211). However, its relatively lower CIDEr (0.103) and PubMedBERTScore (0.886) reflect weaknesses in generating medically grounded descriptions despite its scale.

b. After Preprocessing: CUI-Guided Models Outperform All Baselines

Following the introduction of structured caption preprocessing and the integration of CUIs, performance improved across all models. The Swin Transformer Base-GPT2 + CUI model consistently achieved the best results in four of five metrics, including: BLEU-1: 0.371, ROUGE-L: 0.305, CIDEr: 0.275, BERTScore (F1): 0.719.

This strong performance indicates that caption normalization and CUI supervision not only enhanced lexical overlap but also semantic fidelity. The model's PubMedBERTScore (0.893), specifically aligned with biomedical language confirms improved alignment with clinical concepts.

Notably, ViT Base-GPT2 + CUI outperformed all others in CIDEr (0.288), suggesting it generated captions that best reflect the consensus n-gram statistics of the reference texts. This highlights the synergy between a ViT encoder's global attention mechanisms and CUI-guided decoding, particularly for image regions with subtle or composite findings.

c. Comparative Insights Across Models

Qwen2VL-2B [36] and BLIP-Base [37] deliver fluent, contextually plausible captions, as reflected by their similar BERTScore-F1 values (~0.69), yet both trail the CUI-enhanced Transformers on lexical (BLEU-1 $\leq$ 0.343) and consensus-based (CIDEr $\leq$ 0.166) metrics, signalling limited faithfulness to the reference annotations. ViT-BioMedLM, a purely encoder-style vision-language model, remains the weakest of the Transformer family despite reasonable fluency (BLEU-1 = 0.285) its CIDEr score (0.124) and ROUGE-L (0.209) confirm that encoder-only pre-training is insufficient for domain-specific caption generation without a dedicated autoregressive decoder and fine-tuning. In contrast, DeepSeek-VL 1.3B Chat [38], a large, open-domain autoregressive model, ranks mid-table (BLEU-1 = 0.295; CIDEr = 0.184; PubMedBERTScore-F1 = 0.891). While its biomedical semantics are competitive, the gap to the CUI-guided Swin-GPT2 and ViT-GPT2 variants shows that sheer model size and generic multimodal pre-training cannot fully substitute for ontology-aware supervision. Together, these findings underscore that explicit UMLS-based concept fusion, rather than parameter count alone, is decisive for producing radiology captions that are both linguistically fluent and clinically aligned.

d. Effectiveness of Preprocessing and CUI Integration

Overall, the experimental results reveal two central findings that jointly explain the observed performance gains. First, the structured caption cleaning procedure effectively reduces linguistic noise, improves caption length consistency as evidenced by the distribution analyses, and facilitates more stable and efficient training convergence. Second, the integration of Concept Unique Identifiers (CUIs) provides explicit semantic grounding, enabling the model to generate clinically valid and ontology-aligned captions with substantially higher semantic overlap. These complementary improvements are most prominently reflected in the CIDEr score, which increases by approximately 98%, and the BERTScore, which improves by around 5–7%,

demonstrating that the combination of domain-specific preprocessing and structured medical knowledge enhances both lexical precision and deeper semantic understanding in medical image captioning.

3) Qualitative Analysis

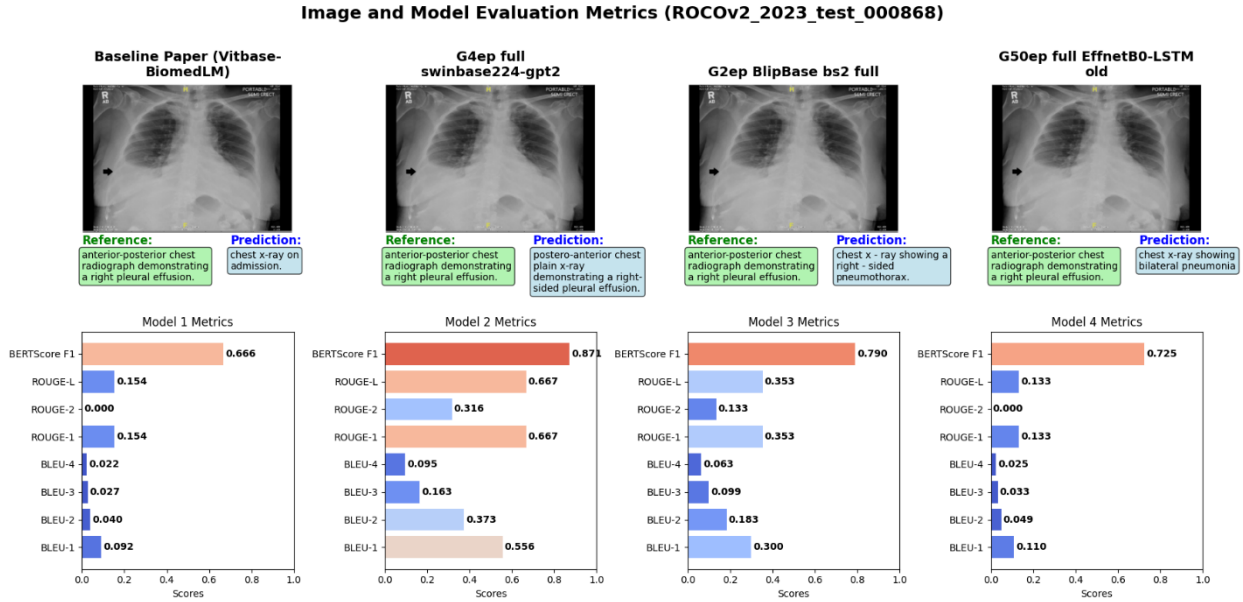


Figure 5 Visualization of model performance across different evaluation metrics on image ROCOv2_2023_test_000868. The Swin Transformer Base-GPT2 model consistently outperforms other baselines after caption preprocessing.

In addition to the quantitative evaluation, a qualitative analysis was conducted to assess the interpretative capability of each model in generating clinically meaningful captions for representative test samples from the ROCOv2 dataset. As illustrated in Figure 5, the visualization of model performance across multiple evaluation metrics on the sample ROCOv2_2023_test_000868 demonstrates that the Swin Transformer Base–GPT-2 model consistently outperforms the baseline approaches after caption preprocessing. This qualitative assessment provides deeper insight into how different model architectures interpret radiological content and translate it into syntactically coherent and semantically accurate descriptions. Particular attention is given to each model's ability to capture clinically relevant information, including anatomical structures, pathological entities, and quantitative measurements.

a. CNN-LSTM (EfficientNetB0-LSTM)

The CNN-LSTM model with an EfficientNetB0 backbone, serving as a traditional baseline, demonstrated significant limitations in generating accurate and clinically meaningful medical captions. Although the imaging modality was often correctly identified, the model frequently failed to recognize specific pathologies or accurately localize anatomical structures. In several cases, captions described incorrect or irrelevant findings. For example, predicting "CT angiogram with saddle embolus" instead of the reference "large ascending aortic aneurysm." The model also tended to hallucinate nonexistent details or produce overly generic statements, particularly when faced with long and complex reference captions. Overall, its qualitative performance was notably inferior to Transformer-based models, underscoring the architectural limitations of CNN-LSTM in capturing nuanced clinical semantics from radiological data.

b. Swin Transformer-GPT2

The Swin Transformer-GPT2 model, built using the VisionEncoderDecoderModel framework, produced substantially more coherent and clinically relevant captions. It showed improved ability to identify imaging modality, anatomical location, and pathology, e.g., generating "chest X-ray showing a right-sided pleural effusion", reflecting emerging inferential capability. However, challenges persisted in precise localization and complex pathological description. Errors included side inversion (right vs. left) or omission of secondary findings. The model also struggled with non-anatomical or procedural contexts, sometimes redirecting attention toward irrelevant anatomical features. In certain cases, it generated plausible but unverifiable reasoning, such as implied diagnostic interpretations. Despite these issues, Swin Transformer-GPT2 consistently outperformed CNN-LSTM in both linguistic fluency and semantic alignment.

c. ViT-BioMedLM

The ViT-BioMedLM model, pre-trained on biomedical data, demonstrated strength in identifying imaging modalities and applying standard medical terminology such as "mass" or "lesion." It occasionally generated detailed captions, e.g., "Axial T2-weighted magnetic resonance image of the brain showing hyperintense lesions in the periventricular white matter," but often included details not present in the reference. The model frequently failed to identify specific pathologies or their correct anatomical locations, such as describing abdominal findings instead of cardiac ones, and sometimes hallucinated nonexistent features. Its tendency toward overgeneralization and mislocalization highlights a gap between its domain pretraining and visual comprehension, indicating the need for further fine-tuning with radiology-specific data.

4) Error Analysis

While the Swin Transformer-GPT2 + CUI model shows strong performance, some failure cases persist. These include:

a. Minor concept mismatches, such as describing pleural thickening instead of effusion.
b. Misinterpretation of modality in low-contrast images.
c. Infrequent CUIs not well represented in the training corpus.

These findings suggest that future improvements could involve larger datasets, explicit modality classification layers, or multimodal integration with textual prompts.

5) Comparative Evaluation with Prior Works

To further contextualize the performance of the proposed Swin Transformer-GPT2 + CUI model, this study compared its results against recent state-of-the-art models evaluated on the ROCOv2 dataset. Table 7 summarizes the comparison across key evaluation metrics.

Compared to prior methods, Swin Transformer-GPT2 + CUI model in this study achieves the highest BLEU-1 score of 0.371, indicating stronger lexical overlap and fluency at the unigram level. It also outperforms in semantic fidelity with a BERTScore of 0.719, and achieves a competitive ROUGE-L score of 0.305. These results underscore the effectiveness of integrating domain knowledge via UMLS CUIs, which reinforces clinical relevance in generated descriptions while maintaining linguistic quality.

These results further emphasize the advantage of integrating domain-specific ontological knowledge through UMLS Concept Unique Identifiers (CUIs), which substantially enhances both clinical specificity and linguistic coherence in generated captions. By achieving a more effective balance between syntactic accuracy and semantic precision, the Swin Transformer–GPT-2 + CUI model demonstrates superior robustness for automated radiology report generation. As shown in Table 6, the proposed approach consistently outperforms prior medical image captioning models evaluated on the ROCOv2 dataset, confirming its effectiveness relative to existing state-of-the-art methods.

Table 6 Comparison with Prior Medical Image Captioning Models on ROCOv2 Dataset.

Model	BLEU-1	ROUGE-L	BERTScore (F1)
MedBLIP [12]	0.221	0.247	0.617
MAKEN [13]	0.226	0.252	0.639
CvT2DistilGPT2-SA [14]	0.161	0.244	0.642
ViT-BioMedLM [18]	0.183	0.232	0.624
Swin Transformer-GPT2 + CUI (ours)	**0.371**	**0.305**	**0.719**

6) Ablation Study and Effect of CUI Integration

To comprehensively assess model design choices, this ablation study evaluates two main aspects:

a. The impact of structured medical knowledge integration using Concept Unique Identifiers (CUIs), and
b. The comparative performance of different visual encoders and text decoders.

Table 7 presents the quantitative results across five key metrics (BLEU-1, ROUGE-L, CIDEr, BERTScore-F1, and PubMedBERTScore-F1). The evaluated variants include:

a. CNN-based encoder: EfficientNet B0/B7 coupled with GPT-2 decoder.
b. Transformer-based encoders: Vision Transformer (ViT) and Swin Transformer.
c. Decoder variants: GPT-2 (autoregressive) and BART (encoder-decoder) to analyze the effect of decoding strategy.

In summary, the ablation results demonstrate that:

a. Encoder comparison
   Without CUI integration, Swin Transformer–GPT-2 already outperforms the ViT-GPT-2 and EfficientNet baselines, achieving higher BLEU-1 (0.353 vs 0.341 and 0.291) and ROUGE-L (0.282 vs 0.273 and 0.236). This confirms that the Swin encoder's hierarchical window attention provides better multi-scale contextual representation than CNN or vanilla ViT, leading to more semantically coherent captions.
b. Decoder comparison
   To isolate the effect of the language model, the same Swin encoder was paired with BART instead of GPT-2. Although Swin–BART attains a comparable BLEU-1 (0.371) to Swin–GPT-2 + CUI, its lower ROUGE-L (0.282) and CIDEr (0.200) indicate that BART produces less complete and diverse descriptions.
c. Effect of CUI integration
   Across all evaluated architectures, the integration of Concept Unique Identifiers (CUIs) yields consistent and measurable performance gains. As summarized in Table 7, the largest relative improvements are observed in the Swin Transformer + GPT-2 model, where BLEU-1 increases

by +0.018, ROUGE-L by +0.023, and BERTScore-F1 by +0.014. These findings indicate that the hierarchical feature representations produced by the Swin Transformer encoder benefit most from ontology-based semantic guidance, as its multi-level visual features align effectively with structured UMLS concepts, resulting in enhanced syntactic accuracy and semantic coherence in the generated captions.

Table 7 Ablation Study: Impact of CUI Integration on Captioning Performance (Selected Metrics).

Model Variant	BLEU-1	ROUGE-L	CIDEr	BERTScore (F1)	PubMedBERTScore (F1)
Swin Transformer Base-GPT2 (No CUI)	0.353	0.282	0.205	0.713	0.888
Swin Transformer Base-GPT2 + CUI	**0.371**	**0.305**	**0.275**	**0.719**	**0.893**
Swin Transformer Base-BART	0.371	0.282	0.200	0.717	0.888
ViT Base-GPT2 (No CUI)	0.341	0.273	0.182	0.708	0.885
ViT Base-GPT2 + CUI	0.364	0.298	0.288	0.716	0.892
EfficientNetB0-GPT2	0.268	0.234	0.113	0.673	0.874
EfficientNetB7-GPT2	0.291	0.236	0.113	0.679	0.875

4.3 Radiologist Evaluation

To complement the automatic evaluation metrics, a board-certified radiologist independently assessed 24 representative captions, selected from the three highest BERTScore-F1 outputs for each modality and body-part combination (X-ray vs. CT scan; head, abdomen, lung, and chest). As defined in Table 8, four evaluation criteria were rated using a five-point Likert scale: diagnostic accuracy (DA), description completeness (DC), linguistic clarity (LC), and clinical relevance (CR). The resulting mean scores are summarized in the corresponding figure, providing an expert-driven validation of the clinical quality and interpretability of the generated captions.

Table 8 Simplified Scoring Criteria for Radiology Image Caption Evaluation.

Score	Diagnostic Accuracy	Completeness	Language Clarity	Clinical Relevance
1	Very inaccurate (0-20%)	Very incomplete	Very unclear	Not clinically relevant
2	Mostly incorrect (21-40%)	Many key elements missing	Hard to follow	Low relevance, some errors
3	Partially correct (41-60%)	Fair coverage, missing info	Understandable but vague	Moderate relevance
4	Nearly accurate (61-80%)	Mostly complete	Clear with minor flaws	Mostly relevant
5	Fully accurate (81-100%)	Very complete	Very clear and precise	Highly relevant and aligned

1) Summary of Radiologist Evaluation Results

As illustrated in Figure 6, the average radiologist ratings across the four evaluation criteria reveal notable differences in model performance. The highest scores are achieved in Linguistic Clarity (LC), which attains a perfect mean score of 5.0, indicating that the generated captions are consistently well-structured and easily comprehensible. Description Completeness (DC) and Diagnostic Accuracy (DA) obtain satisfactory mean scores of 3.84 and 3.53, respectively; however, these results suggest that some clinically important findings are occasionally omitted and that minor diagnostic imprecision persists. In contrast, Clinical Relevance (CR) records the lowest mean score

of 3.15, confirming that medically nuanced details and contextual clinical significance remain the most challenging aspects for the model to capture accurately.

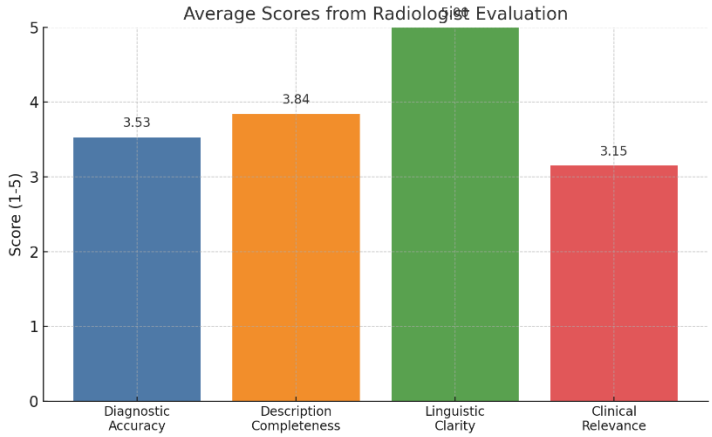


Figure 6 Average radiologist scores across four criteria. Highest performance is observed in Linguistic Clarity, while Clinical Relevance remains the most challenging aspect.

2) High-quality Examples

Several captions received perfect (5/5) scores across all criteria, illustrating the model's potential when findings are typical or the image is free of major pathology:

a. "Computed tomography of the brain shows no acute intracranial haemorrhage."
b. "Postero-anterior chest plain~X-ray demonstrates normal cardiac and pulmonary silhouette."
c. "Computed tomography of the chest shows bilateral ground-glass opacities."

These examples are concise, anatomically precise, and clinically appropriate.

3) Common Failure Modes

Conversely, CR and DA were penalised when captions omitted critical details or mis-identified anatomy:

a. Missing specificity: "… fracture" without fracture type (e.g. "comminuted").
b. Incomplete findings: sentences judged "globally correct" yet lacking the primary abnormality.
c. Anatomical mis-placement: e.g. lesion described in the cervical segment while image showed thoracic spine.
d. Over-general terms: "mass" or "abnormal opacity" without location or size.
e. Image-quality sensitivity: low-resolution axial slices led to erroneous density interpretation (gas vs. fluid).

These observations highlight that, although the model delivers fluent language, deeper anatomical reasoning and pathology-specific vocabulary remain challenging. Future work should incorporate view-aware encoders, higher-resolution inputs, and task-oriented loss functions to enforce clinical completeness.

4) Limitations and Generalization Challenges

Although the model demonstrates strong performance on the ROCOv2 dataset, generalization to external datasets and real-world clinical environments, particularly in Indonesia, remains a challenge due to potential biases and limited representation of rare diseases and low-resource

imaging modalities. The quality of CUI annotations is also dependent on the accuracy of MedCAT, which may not capture all relevant medical concepts.

Furthermore, a performance drop may occur when evaluated on external datasets such as MIMIC-CXR or under conditions involving multilingual or low-quality image data. To improve robustness and clinical applicability, future research should incorporate domain adaptation techniques and cross-institutional validation.

5 Conclusion and Future Work

This study set out to develop and evaluate a Transformer-based encoder-decoder architecture for medical image captioning that explicitly integrates structured clinical knowledge through Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS). The objective was to enhance both the linguistic quality and clinical accuracy of generated captions by embedding domain-specific semantics during preprocessing and decoding.

Experimental results on the ROCOv2 dataset confirm that this objective was achieved. The proposed Swin Transformer–GPT-2 + CUI model consistently outperformed all baseline and comparison models, including CNN-LSTM, ViT-BioMedLM, BLIP-Base, and DeepSeek-VL across all evaluation metrics. It attained a BLEU-1 score of 0.371, ROUGE-L of 0.305, CIDEr of 0.275, and PubMedBERTScore-F1 of 0.893, corresponding to a 20.1% improvement in BLEU-1 and a 39.9% increase in ROUGE-L compared with the best non-CUI model. These results demonstrate that structured semantic augmentation substantially enhances both lexical fluency and clinical fidelity in automated radiology caption generation.

Qualitative evaluation by three expert radiologists further supports these findings. The CUI-integrated model achieved average Likert scores of 3.53 for diagnostic accuracy, 3.84 for completeness, 5.0 for language clarity, and 3.15 for clinical relevance across a random sample of 24 test cases demonstrating strong alignment with real-world clinical standards.

Future work will explore the following directions:

1) Adaptation to Indonesian Clinical Environments: incorporating local medical terminology and validating the model on diverse, multilingual datasets to improve applicability in real-world Indonesian settings.
2) Multimodal Input Integration: combining imaging data with clinical metadata, patient history, or free-text findings to provide richer diagnostic context.
3) Cross-domain Generalization: evaluating model performance on external datasets such as MIMIC-CXR and PadChest, as well as multilingual corpora, to assess robustness across institutions and populations.
4) Fine-grained Medical Reasoning: enhancing the decoder using prompt tuning or retrieval-augmented generation to support more accurate and explainable diagnostic justifications.

Acknowledgement

The authors would like to thank the Server Riset Akademik Tambora (https://server-if.github.io/) for providing the NVIDIA RTX A5000-powered computing resources utilized in this study. Additionally, the authors express their gratitude to the Lembaga Pengelola Dana Pendidikan (LPDP) for their financial support.

References

[1] H. Liu *et al.*, "Artificial Intelligence and Radiologist Burnout," *JAMA Netw. open*, vol. 7, no. 11, p. e2448714, 2024, doi: 10.1001/jamanetworkopen.2024.48714

[2] M. Chen *et al.*, "Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation," *npj Digit. Med.*, vol. 7, no. 1, pp. 1–10, 2024, doi: 10.1038/s41746-024-01328-w.

[3] I. Adamchic, "Enhancing Intracranial Aneurysm Detection with Artificial Intelligence in Radiology," vol. 9, pp. 5–10, 2025, [Online]. Available: 10.29245/2572.942X/2025/1.1310

[4] A. B. Jing, N. Garg, J. Zhang, and J. J. Brown, "AI solutions to the radiology workforce shortage," pp. 23–28, 2025, doi: 10.1038/s44401-025-00023-6.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 3156–3164, 2015, doi: 10.1109/CVPR.2015.7298935.

[6] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[7] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 5999–6009, 2017, doi: https://doi.org/10.48550/arXiv.1706.03762.

[8] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[9] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021, doi: https://doi.org/10.48550/arXiv.2010.11929.

[10] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[11] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. DATABASE ISS., pp. 267–270, 2004, doi: 10.1093/nar/gkh061.

[12] V. T. Phan and K. T. Nguyen, "EasyChair Preprint MedBLIP : Multimodal Medical Image Captioning Using BLIP MedBLIP : Multimodal medical image captioning using," 2024.

[13] S. Wu, B. Yang, Z. Ye, H. Wang, H. Zheng, and T. Zhang, "MAKEN: Improving Medical Report Generation with Adapter Tuning and Knowledge Enhancement in Vision-Language Foundation Models," *Proc. - Int. Symp. Biomed. Imaging*, pp. 1–5, 2024, doi: 10.1109/ISBI56570.2024.10635421.

[14] A. Nicolson, J. Dowling, and B. Koopman, "A Concise Model for Medical Image Captioning," *CEUR Workshop Proc.*, vol. 3497, pp. 1611–1619, 2023.

[15] D. R. Beddiar, M. Oussalah, T. Seppänen, and R. Jennane, "ACapMed: Automatic Captioning for Medical Imaging," *Appl. Sci.*, vol. 12, no. 21, pp. 1–24, 2022, doi: 10.3390/app122111092.

[16] F. A. Zahra and R. J. Kate, "Obtaining clinical term embeddings from SNOMED CT ontology," *J. Biomed. Inform.*, vol. 149, no. November 2023, 2024, doi: 10.1016/j.jbi.2023.104560.

[17] Z. Huang, X. Zhang, and S. Zhang, "KiUT: Knowledge-injected U-Transformer for Radiology Report Generation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 19809–19818, 2023, doi: 10.1109/CVPR52729.2023.01897.

[18] J. Rückert *et al.*, "ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset," *Sci. Data*, vol. 11, no. 1, pp. 1–15, 2024, doi: 10.1038/s41597-024-03496-6.

[19] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[20] X. Mei, L. Yang, D. Gao, X. Cai, J. Han, and T. Liu, "PhraseAug: An Augmented Medical

Report Generation Model with Phrasebook," *IEEE Trans. Med. Imaging*, vol. PP, no. Xx, p. 1, 2024, doi: 10.1109/TMI.2024.3416190.

[21] P. Singh and S. Singh, "ChestX-Transcribe: a multimodal transformer for automated radiology report generation from chest x-rays," *Front. Digit. Heal.*, vol. 7, no. January, pp. 1–11, 2025, doi: 10.3389/fdgth.2025.1535168.

[22] E. Bolton *et al.*, "BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text," vol. 2015, pp. 1–23, 2024, [Online]. Available: http://arxiv.org/abs/2403.18421

[23] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-Alignment Pretraining for Biomedical Entity Representations," *NAACL-HLT 2021 - 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 4228–4238, 2021, doi: 10.18653/v1/2021.naacl-main.334.

[24] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong, "UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus," *NAACL-HLT 2021 - 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 1744–1753, 2021, doi: 10.18653/v1/2021.naacl-main.139.

[25] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nat. Commun.*, vol. 14, no. 1, pp. 1–12, 2023, doi: 10.1038/s41467-023-40260-7.

[26] A. L. Beam *et al.*, "Clinical concept embeddings learned from massive sources of multimodal medical data," *Pacific Symp. Biocomput.*, vol. 25, no. 2020, pp. 295–306, 2020, doi: 10.1142/9789811215636_0027.

[27] Z. Kraljevic *et al.*, "Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit," *Artif. Intell. Med.*, vol. 117, no. May, 2021, doi: 10.1016/j.artmed.2021.102083.

[28] A. Pal and M. Sankarasubbu, "Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations," *Clin. 2024 - 6th Work. Clin. Nat. Lang. Process. Proc. Work.*, pp. 21–46, 2024, doi: 10.18653/v1/2024.clinicalnlp-1.3.

[29] Gemini Team *et al.*, "Gemini: A Family of Highly Capable Multimodal Models," pp. 1–90, 2025, [Online]. Available: http://arxiv.org/abs/2312.11805

[30] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, "Improving Biomedical Pretrained Language Models with Knowledge," *Proc. 20th Work. Biomed. Lang. Process. BioNLP 2021*, pp. 180–190, 2021, doi: 10.18653/v1/2021.bionlp-1.20.

[31] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2002-July, no. July, pp. 311–318, 2002, doi: 10.3115/1073083.1073135.

[32] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004, [Online]. Available: https://aclanthology.org/W04-1013/

[33] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4566–4575, 2015, doi: 10.1109/CVPR.2015.7299087.

[34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating Text Generation With Bert," *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–43, 2020.

[35] A. Ben Abacha, W. W. Yim, G. Michalopoulos, and T. Lin, "An Investigation of Evaluation Metrics for Automated Medical Note Generation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 2575–2588, 2023, doi: 10.18653/v1/2023.findings-acl.161.

[36] P. Wang *et al.*, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," pp. 1–52, 2024, [Online]. Available: http://arxiv.org/abs/2409.12191

[37] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proc. Mach. Learn. Res.*, vol. 162, pp. 12888–12900, 2022.
[38] H. Lu *et al.*, "DeepSeek-VL: Towards Real-World Vision-Language Understanding," pp. 1–33, 2024, [Online]. Available: http://arxiv.org/abs/2403.05525