DS-MED Lab at ImageCLEFmedical Caption 2026: Clinical-Aware Routing for Hallucination Suppression and Pathology-Focused Medical Image Captioning

Notebook for the DS-MED Lab (UIT) at CLEF 2026

Tien-Thinh Nguyen-Chi[1,2], Bao Quoc Pham Ho[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents our system for the ImageCLEFmedical Caption 2026 task on automated medical image captioning using the ROCOv2 radiology dataset. Preliminary analysis of a vanilla BLIP-large baseline revealed two clinically undesirable behaviors: frequent anatomical hallucination and omission of relevant pathological findings, highlighting the limitations of optimizing caption generation primarily through lexical similarity objectives. To address these issues, a clinical-aware framework is proposed, prompt-guided framework that incorporates explicit semantic priors before autoregressive caption generation. The proposed system consists of three components. First, a multimodal semantic clustering module based on BiomedCLIP partitions the dataset into coarse clinical macro-domains using K-Means clustering over joint image-text embeddings. Second, a dual-branch EfficientNet-B4 classifier enhanced with Class-Specific Residual Attention (CSRA) and trained using Asymmetric Loss performs multi-label Concept Unique Identifier (CUI) detection. Third, the predicted CUIs and domain tokens are integrated into a cluster-aware prompting mechanism that conditions a fine-tuned BLIP-large decoder through Scheduled Teacher Forcing Dropout. Internal validation on the ROCOv2 validation set shows that the proposed framework reduces anatomical hallucination from 59.4% to 39.2% and improves Pathology Recall from 0.3583 to 0.5312 compared with the vanilla BLIP-large baseline. On the official ImageCLEFmedical Caption 2026 test set, the submitted system achieved an Overall score of 0.2814 and a Similarity score of 0.7286. Although the official factuality-related metrics remain challenging, our results suggest that concept-aware prompting can improve pathology-oriented grounding while reducing anatomically inconsistent generation. More broadly, the findings highlight the trade-off between conventional lexical evaluation metrics and clinically meaningful caption generation in radiology vision-language systems.

Keywords

ImageCLEF 2026, Medical Image Captioning, Hallucination Suppression, Clinical Concept Detection, BLIP, EfficientNet-B4, CSRA, BiomedCLIP, Semantic Clustering, Prompt-Guided Generation, Radiology Images, Vision-Language Models, Asymmetric Loss, ROCOv2

1. Introduction

Recent advances in deep learning and vision-language modeling have major improved automated biomedical image understanding. In radiology, medical image captioning aims to generate clinically meaningful textual descriptions from medical images such as X-rays, CT scans, and MRI scans. Such systems have the potential to assist clinical workflows by supporting report drafting, improving information accessibility, and highlighting visually salient findings.

The ImageCLEFmedical Caption task [1], organized within the broader ImageCLEF evaluation campaign [2], provides a standardized benchmark for studying this problem on the ROCOv2 dataset [3]. Participants are required to generate medically relevant captions for radiology images while simultaneously addressing concept detection and semantic consistency challenges.

Recent progress in medical image captioning has been largely driven by pretrained Vision-Language Models (VLMs), particularly BLIP-based architectures [4]. These models can achieve strong performance

CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany
*Corresponding author.
24521688@gm.uit.edu.vn (T. Nguyen-Chi); 24520152@gm.uit.edu.vn (B. Q. P. Ho); thienntb@uit.edu.vn (T. B. Nguyen-Tat)
0009-0005-6609-4753 (T. Nguyen-Chi); 0009-0002-7010-844X (B. Q. P. Ho); 0000-0002-4809-7126 (T. B. Nguyen-Tat)

on lexical and semantic similarity metrics such as BERTScore [5]. However, high similarity scores do not necessarily imply clinical correctness. In practice, VLMs may generate fluent but clinically inconsistent descriptions, especially when the visual evidence is ambiguous or weakly grounded.

One important failure mode is *anatomical hallucination*, where the model generates anatomical structures that are not present in the image. Another is *pathological omission*, where clinically relevant abnormalities are not mentioned in the generated caption [6]. In medical settings, such errors are particularly problematic because semantically plausible captions may still omit diagnostically important findings.

To better understand these issues, we conducted a preliminary analysis using a vanilla BLIP-large baseline on the ROCOv2 validation set. The analysis revealed frequent anatomical hallucination together with limited pathology recall, suggesting that similarity-oriented optimization alone is insufficient for clinically grounded caption generation. These observations motivated the design of a concept-aware captioning framework that explicitly incorporates semantic priors before autoregressive decoding.

In this work, the proposed architecture is a clinical-aware, prompt-guided architecture for the ImageCLEFmedical Caption 2026 task. Motivated by prior ImageCLEFmedical approaches integrating concept detection and caption generation [7], the proposed framework integrates three main components:

1. A multimodal semantic clustering module based on BiomedCLIP [8] to derive coarse clinical macro-domains;
2. A dual-branch EfficientNet-B4 [9] concept detector enhanced with Class-Specific Residual Attention (CSRA) [12] and Asymmetric Loss [13] for multi-label CUI prediction;
3. A cluster-aware prompting mechanism that conditions a fine-tuned BLIP-large [4] decoder on predicted concepts and routing tokens.

Rather than focusing solely on improving lexical overlap with reference captions, our framework is designed to encourage pathology-oriented grounding while reducing anatomically inconsistent generation. Internal validation on the ROCOv2 validation set shows that the proposed system reduces anatomical hallucination and improves pathology recall compared with a vanilla BLIP-large baseline, although challenges in factual grounding remain under the official evaluation metrics.

The remainder of this paper is organized as follows. Section 2 reviews related work in medical image captioning and hallucination-aware vision-language modeling. Section 3 describes the dataset and task setup. Section 4 presents the proposed framework. Section 5 reports experimental results and internal clinical validation. Section 6 discusses the trade-offs between lexical similarity and clinical grounding, and Section 7 outlines future research directions.

2. Related Work

2.1. ImageCLEFmedical Caption Task

The ImageCLEFmedical Caption task [1] has evolved into a widely used benchmark for medical concept detection and radiology image caption generation. Recent editions have progressively expanded both the scale of the dataset and the scope of evaluation. The 2024 edition [14] introduced the ROCOv2 dataset [3] and evaluated systems on concept detection and caption prediction subtasks. The 2025 edition [15] further incorporated interpretable caption generation, emphasizing the importance of explainability and clinical consistency in medical vision-language systems.

These benchmark campaigns have highlighted the growing adoption of pretrained Vision-Language Models (VLMs) together with concept-aware pipelines for radiology captioning. At the same time, the shared tasks continue to expose challenges related to semantic grounding, factual consistency, and clinically meaningful evaluation.

2.2. Medical Image Captioning with Vision-Language Models

Early medical image captioning approaches commonly relied on convolutional visual encoders coupled with recurrent language decoders. More recent systems increasingly adopt Transformer-based Vision-

Language Models pretrained on large-scale image-text corpora. Among them, BLIP [4] has emerged as a strong foundation for medical caption generation due to its unified image-text pretraining objective and flexible generative capabilities.

Several ImageCLEFmedical participants have explored BLIP-based architectures for radiology captioning. Phan et al. [16] fine-tuned BLIP on the ROCOv2 dataset and reported competitive semantic similarity performance. Nguyen et al. [17] investigated Transformer-based captioning systems for radiology images, emphasizing the role of domain-adapted visual representations. More recently, Pan et al. [18] combined a fine-tuned vision-language model with SapBERT-based reranking strategies for concept-aware generation, reflecting a broader trend toward integrating structured clinical semantics into generative pipelines.

Despite these advances, existing systems are still largely optimized using lexical or semantic similarity objectives, which may not adequately capture clinical correctness or pathological completeness.

2.3. Concept Detection and Multi-Label Medical Classification

Medical concept detection is typically formulated as a multi-label classification problem, where the model predicts a set of UMLS Concept Unique Identifier (CUIs) associated with an image. In radiology settings, this task is particularly challenging due to severe class imbalance and the localized nature of many pathological findings.

Recent studies have explored the integration of concept detection into caption generation pipelines. Nguyen et al. [19] demonstrated that incorporating detected concepts as auxiliary conditioning signals can improve concept coverage during caption generation. Similarly, prior ImageCLEFmedical systems [7] investigated CSRA-enhanced concept extraction combined with BLIP-based captioning frameworks.

For multi-label recognition, EfficientNet [9] remains a widely adopted visual backbone because of its favorable accuracy-efficiency trade-off. In addition, Class-Specific Residual Attention (CSRA) [12] has shown effectiveness in localizing discriminative regions for multi-label classification, while Asymmetric Loss (ASL) [13] provides improved robustness under long-tail label distributions.

2.4. Hallucination in Vision-Language Models

Hallucination has been identified as a major limitation of modern Vision-Language Models. Rohrbach et al. [6] showed that image captioning systems frequently generate objects that are not visually grounded in the input image, even when the generated captions remain linguistically plausible.

In medical applications, hallucination can be particularly problematic because generated anatomical structures or findings may not correspond to actual clinical evidence. Recent ImageCLEFmedical evaluations [15] have also highlighted the importance of factual consistency and explainability in radiology caption generation. However, most existing approaches continue to prioritize lexical similarity metrics, while explicit analysis of anatomical hallucination and pathological omission remains relatively limited.

Our work builds upon these observations by explicitly evaluating hallucination and omission behaviors using a dual-concept diagnostic dictionary covering anatomical structures and pathological findings.

2.5. Prompt-Guided and Concept-Aware Generation

Prompt-guided generation has recently emerged as an effective strategy for adapting pretrained language models to domain-specific tasks. In medical vision-language systems, prompts can provide structured semantic priors that guide autoregressive decoding toward clinically relevant outputs.

Several recent ImageCLEFmedical systems incorporate detected concepts or retrieved semantic evidence into the generation process [18, 19]. Our approach follows this direction by integrating both fine-grained CUI predictions and coarse-grained semantic routing tokens derived from BiomedCLIP [8] clustering. This design aims to provide complementary local and global contextual guidance for the BLIP decoder during caption generation.

3. Dataset

3.1. ROCOv2 Dataset

Our experiments were conducted using an updated and extended version of the ROCOv2 dataset [3] provided as the official benchmark of the ImageCLEFmedical Caption 2026 task [1]. ROCOv2 is a large-scale multimodal radiology dataset constructed from biomedical articles within the PubMed Central OpenAccess collection. Each sample consists of a medical image, its associated figure caption, and a set of UMLS concepts manually curated as structured annotations.

The dataset contains diverse radiological images, covering a broad range of anatomical regions and pathological conditions. Following the official protocol, the dataset is divided into training, validation, and hidden test splits for model development and benchmark evaluation.

3.2. Task Description

We participated in the two primary subtasks of ImageCLEFmedical Caption 2026:

Concept Detection. Given a medical image, the objective is to predict a set of clinically relevant UMLS Concept Unique Identifier (CUIs). The label space is derived from the UMLS 2022 AB release after semantic filtering and frequency-based pruning. Official evaluation is performed using concept-level F_1-score.

Caption Prediction. Given a medical image, the system generates a clinically relevant textual description. The official ranking is computed using a composite score combining lexical, semantic, and factuality-oriented metrics. These include BERTScore [5], ROUGE-1 [10], BLEURT [20], image-text similarity [1], MedCAT-based concept overlap [21], and AlignScore-based factual consistency evaluation [11].

In the 2026 edition, the Overall score serves as the primary ranking metric, with BERTScore and ROUGE-1 used as secondary criteria. Prior to evaluation, captions are normalized by lowercasing, punctuation removal, and numerical token standardization according to the official evaluation pipeline.

3.3. Data Preprocessing

For caption generation, hyperlinks and formatting artifacts were removed from the training captions. Images were resized and normalized according to the preprocessing requirements of the corresponding visual backbone models.

For concept detection, infrequent CUIs were filtered to reduce extreme label sparsity and improve training stability in the multi-label classification setting. All experiments strictly followed the official task constraints, without mixing standard and synthetic task subsets.

4. Proposed Method

4.1. Motivation and System Overview

Although recent Vision-Language Models (VLMs) achieve strong performance on general captioning benchmarks, preliminary experiments on the ROCOv2 validation set revealed that a vanilla BLIP-large model frequently produced clinically inconsistent descriptions. In particular, two recurrent behaviors were identified:

1. **Anatomical Hallucination:** generation of anatomical structures not supported by the image content;
2. **Pathological Omission:** failure to mention clinically relevant pathological findings present in the reference caption.

To quantify these behaviors, a dual-concept diagnostic dictionary was constructed diagnostic dictionary consisting of two semantic categories: *Anatomical Structures* and *Pathological Findings*. Using

this protocol, the baseline BLIP-large model exhibited an anatomical hallucination rate of 59.4% and limited pathology recall on the validation set. These observations suggest that optimization toward lexical similarity alone may not sufficiently encourage clinically grounded caption generation.

Motivated by this limitation, a clinical-aware framework is introduced, prompt-guided framework that incorporates explicit semantic priors before autoregressive decoding. The proposed architecture combines:

- fine-grained clinical concepts represented by predicted UMLS Concept Unique Identifiers (CUIs);
- and coarse-grained semantic routing information derived from multimodal clustering.

The overall pipeline consists of three stages:

1. **Semantic Clustering and Routing:** multimodal semantic clustering together with a visual routing network for macro-domain prediction;
2. **Medical Concept Extraction:** multi-label CUI prediction using an EfficientNetCSRA classifier;
3. **Cluster-Aware Caption Generation:** prompt-guided BLIP-large decoding conditioned on predicted concepts and routing tokens.

4.2. Overall Architecture

Figure 1 illustrates the overall architecture of the proposed framework. Given an input image I, the system first predicts a semantic routing label and a set of clinical concepts. These predictions are then converted into textual prompts that condition the BLIP-large decoder during caption generation.

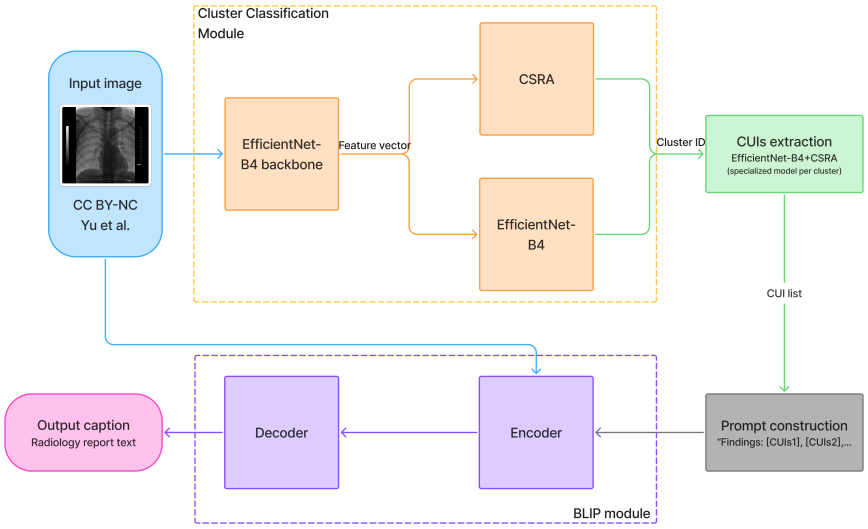

Figure 1: The overall architecture of the proposed framework, illustrating the integration of semantic cluster classification, clinical concept extraction, and prompt-conditioned caption generation.

Unlike conventional end-to-end captioning systems that rely solely on implicit visual-language alignment, our framework introduces an intermediate semantic guidance stage to improve concept grounding and reduce clinically inconsistent generations.

4.3. Problem Formulation

Let I denote the input image and $Y = \{y_1, y_2, \ldots, y_T\}$ denote the target caption sequence. To incorporate semantic guidance during generation, we introduce:

- a fine-grained concept variable $C = \{c_1, c_2, \ldots, c_K\}$ representing predicted CUIs;

- and a coarse-grained routing variable $Z \in \{1, 2, 3\}$ representing semantic macro-domains.

The proposed framework can be conceptually decomposed as:

$$P(Y, C, Z \mid I) = P(Y \mid I, C, Z) \cdot P(C \mid I) \cdot P(Z \mid I) \tag{1}$$

where, in practice, the full distributions $P(C \mid I)$ and $P(Z \mid I)$ are approximated by point estimates:

$$\hat{C} = f_{\text{CUI}}(I), \qquad \hat{Z} = f_{\text{router}}(I) \tag{2}$$

where f_{CUI} and f_{router} denote the deterministic outputs of the concept extractor and routing network respectively, rather than full posterior distributions.

The predicted concepts and routing labels are transformed into a textual prompt P that conditions the autoregressive decoder. The captioning module, parameterized by $\boldsymbol{\theta}$, is optimized using the standard negative log-likelihood objective:

$$\mathcal{L}_{\text{caption}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I, C, Z; \boldsymbol{\theta}) \tag{3}$$

Although the proposed semantic guidance mechanism improves concept grounding, it does not guarantee factual correctness, as the final caption quality still depends on both the reliability of predicted CUIs and the generative behavior of the decoder.

4.4. Semantic Clustering Module

To provide coarse-grained semantic guidance, we organize the training samples into multimodal macro-domains using BiomedCLIP [8] (`PubMedBERT_256-vit_base_patch16_224`). Figure 2 illustrates the overall pipeline of this module.

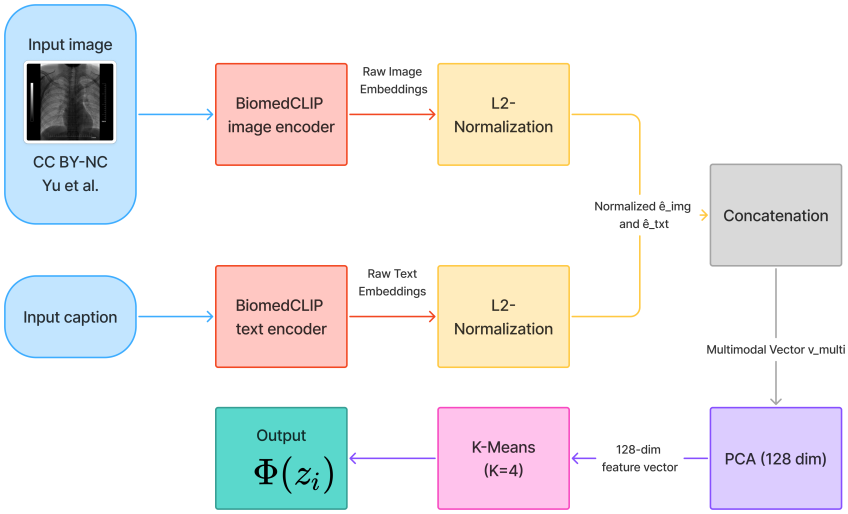

Figure 2: Overview of the cluster classification module. The process leverages pre-trained BiomedCLIP encoders to form a joint multimodal space, followed by PCA dimensionality reduction and K-Means clustering to determine the routing variable.

4.4.1. Multimodal Feature Extraction

For each training sample, the image I and corresponding caption T are independently encoded using BiomedCLIP. The resulting embeddings are L_2-normalized and concatenated:

$$\mathbf{v}_{\text{multi}} = \left[\hat{\mathbf{e}}_{\text{img}} \,\|\, \hat{\mathbf{e}}_{\text{txt}}\right] \tag{4}$$

To reduce redundancy and high-dimensional noise, Principal Component Analysis (PCA) [22] is applied to project the representation into a 128-dimensional subspace:

$$\mathbf{z} = \mathbf{W}_{\text{PCA}}^{T}(\mathbf{v}_{\text{multi}} - \boldsymbol{\mu}) \tag{5}$$

4.4.2. K-Means Clustering and Consolidation

We apply K-Means clustering [23] with $K = 4$ on the compressed feature space. During empirical inspection, two clusters exhibited highly overlapping anatomical distributions and visually similar layouts. To simplify the downstream routing task, these clusters were merged into a single category. The final mapping function $\Phi(\cdot)$ produces three semantic macro-domains:

$$\Phi(\mathbf{z}_i) = \begin{cases} 1 & \text{if } \mathbf{z}_i \in \mathcal{C}_1 \\ 2 & \text{if } \mathbf{z}_i \in \mathcal{C}_2 \\ 3 & \text{if } \mathbf{z}_i \in (\mathcal{C}_3 \cup \mathcal{C}_4) \end{cases} \tag{6}$$

Figure 3 presents the 2D PCA projection of the multimodal feature space along with the per-cluster image distribution, while Figure 4 illustrates the consolidated clustering results after merging the two overlapping domains. These cluster assignments are subsequently used as supervisory targets for the routing network.

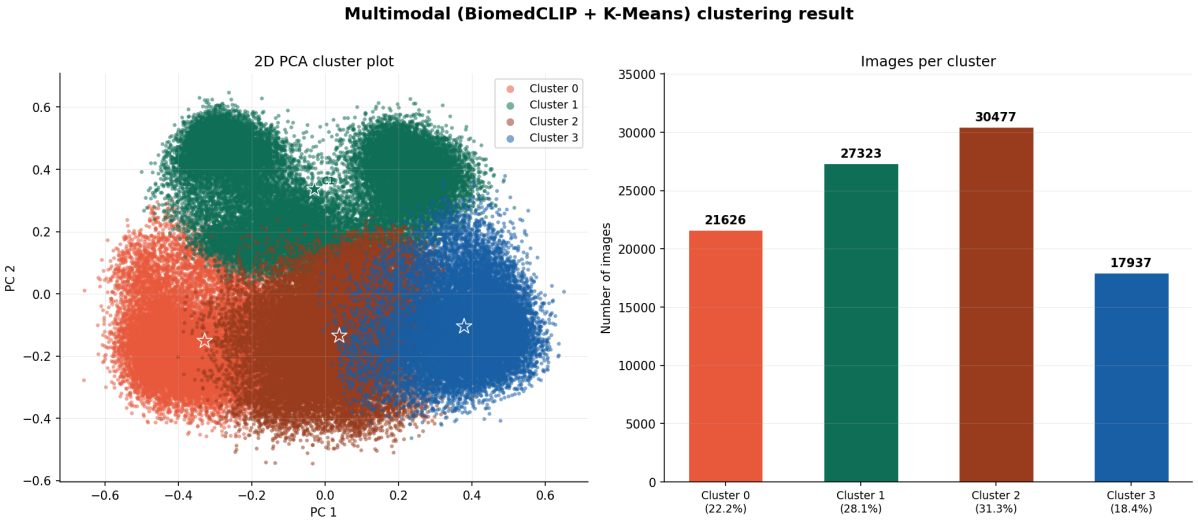

Figure 3: 2D PCA projection and data distribution of the multimodal feature space clustered using K-Means ($K = 4$). The left panel shows the spatial boundaries and centroids, while the right panel displays the number of images per cluster.

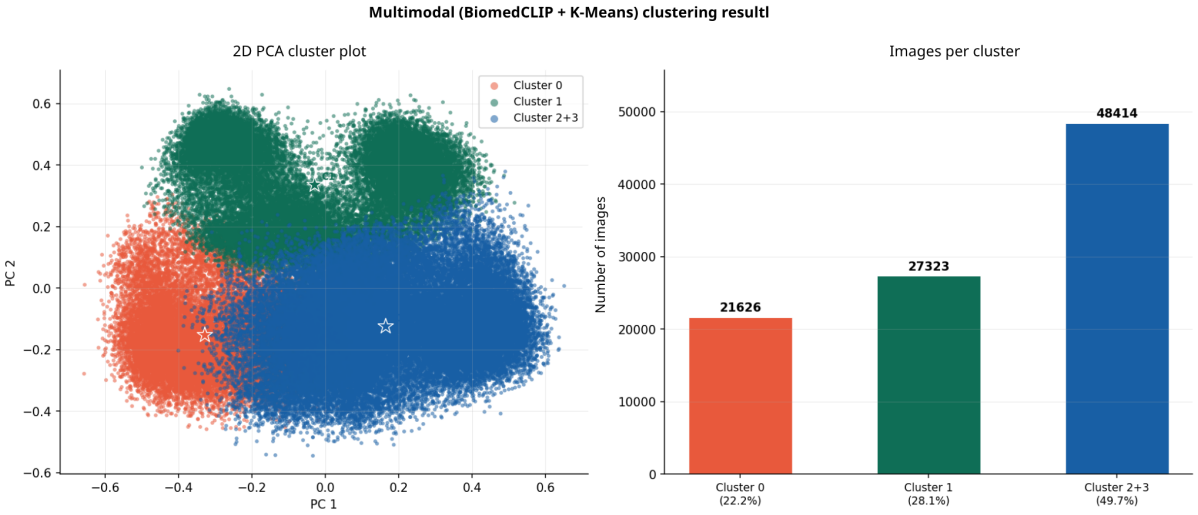

Figure 4: Consolidated clustering results mapping to three semantic macro-domains. Two highly overlapping categories are merged into a single domain ($\mathcal{C}_3 \cup \mathcal{C}_4$) to simplify the downstream routing task.

4.5. Medical Concept Extraction Module

To predict clinical concepts from medical images, we implemented a multi-label classification module based on EfficientNet-B4 [9] enhanced with Class-Specific Residual Attention (CSRA) [12]. The resulting architecture is denoted as `EfficientNetCSRA` and is depicted in Figure 5.

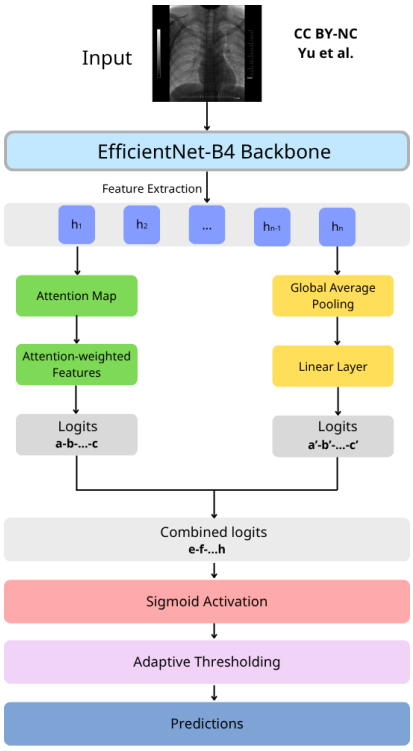

Figure 5: Overview of the `EfficientNetCSRA` architecture. The model leverages an EfficientNet-B4 backbone for feature extraction, followed by a dual-branch CSRA module that combines attention-weighted features and global average pooled features to predict multi-label clinical concepts.

4.5.1. Two-Phase Optimization and Feature Fusion

Training is performed in two stages:

1. freezing the backbone and optimizing only the classification heads for 20 epochs;
2. partially unfreezing the backbone from block 6 onward for an additional 10 epochs.

Given a spatial feature tensor $\mathbf{F} \in \mathbb{R}^{B\times D\times H\times W}$, two complementary feature branches are applied:

- a Global Average Pooling branch producing $\mathbf{z}_{\text{global}}$;
- and a CSRA branch producing $\mathbf{z}_{\text{csra}}$.

The final logits are computed using weighted fusion:

$$\mathbf{z} = (1-\lambda)\mathbf{z}_{\text{global}} + \lambda\mathbf{z}_{\text{csra}} \tag{7}$$

where $\lambda \in [0,1]$ is a fixed fusion weight controlling the relative contribution of the CSRA branch. In all experiments, λ was set to 0.1 following the default configuration of the original CSRA implementation [12].

4.5.2. Asymmetric Loss and Adaptive Thresholding

Because the CUI distribution is highly imbalanced, Asymmetric Loss is employed (ASL) [13] with $\Delta = 0.05$, $\gamma_{\text{pos}} = 1.0$, and $\gamma_{\text{neg}} = 4.0$.

Instead of using a fixed prediction threshold, we determine a class-specific threshold τ_c on the validation set:

$$\tau_c = \arg\max_{th\in\mathcal{T}} F_1\left(\mathbb{I}(\sigma(z_{ic}) \geq th), y_{ic}\right) \tag{8}$$

This adaptive thresholding strategy improves robustness under long-tail label distributions.

4.5.3. Visual Routing Network

Since textual metadata is unavailable during inference, the semantic routing labels must be predicted directly from the image. A dedicated visual routing network is therefore trained using the same `EfficientNetCSRA` architecture, producing the routing variable:

$$Z = \hat{y}_{\text{cluster}} \in \{1,2,3\} \tag{9}$$

using Cross-Entropy supervision over the consolidated cluster assignments generated by $\Phi(\cdot)$.

The routing variable Z and the concept set C are predicted independently by two separate networks with no shared parameters. This decoupled design was adopted primarily to simplify the training pipeline, allowing each module to be developed and optimized in isolation without requiring joint supervision.

It is worth noting that the cluster assignments used as training targets are derived from joint image-text embeddings produced by BiomedCLIP, whereas the routing network approximates these multimodal cluster boundaries using visual information alone. This approximation is necessary given that textual captions are unavailable at inference time.

4.6. Cluster-Aware Prompting and BLIP Generation

Figure 6 illustrates the architectural pipeline of the BLIP generation module, showing how visual and textual concept features are integrated to condition text generation.

4.6.1. Image Preprocessing and Concept Translation

Input images are first enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) [24] with $\alpha = 2.0$ on an 8×8 grid configuration, then converted into pseudo-RGB format. Predicted CUIs are subsequently mapped into canonical medical terminology through a deterministic lexical mapping function Ψ, producing the concept token sequence $\mathcal{T}_{\text{findings}}$.

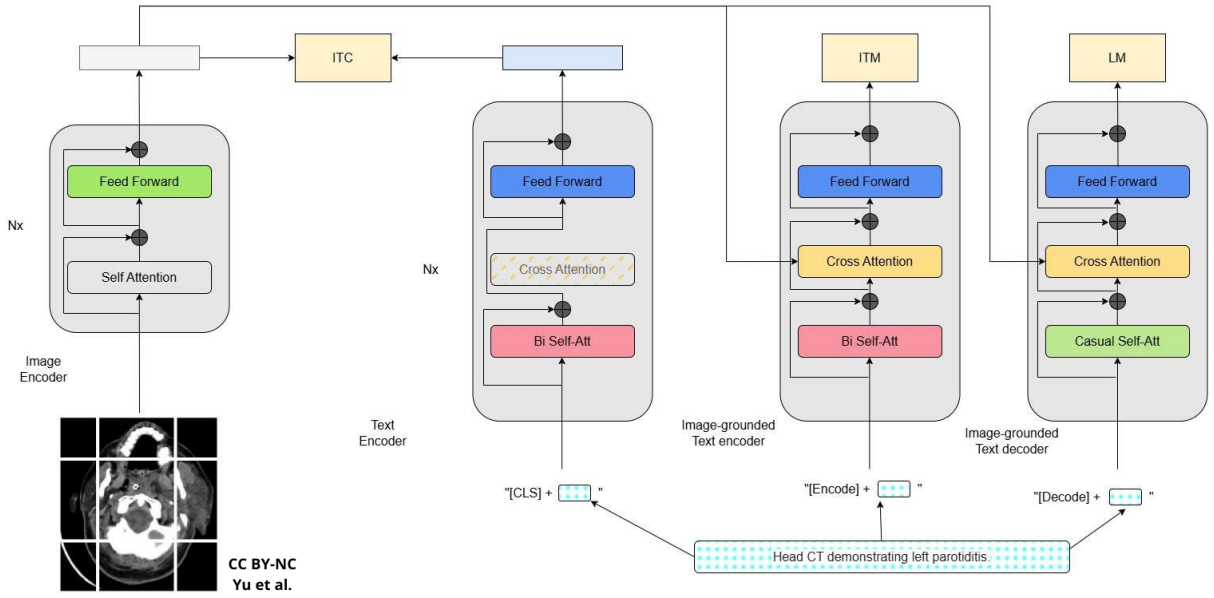

Figure 6: Architectural pipeline of the BLIP generation module, illustrating the detailed multi-modal attention mechanisms used to condition text generation on visual and textual concept features.

4.6.2. Prompt Construction and Dynamic Prompt Fading

The final prompt P combines:

- a fixed system instruction;
- predicted medical findings;
- and semantic routing tokens.

To reduce dependence on ground-truth prompts during training, a Scheduled Teacher Forcing Dropout strategy is adopted. The probability of retaining the findings prompt decreases linearly throughout training:

$$p_{\text{keep}} = \max\left(0.0, 1.0 - \frac{\lfloor e/2 \rfloor}{\lfloor (E-1)/2 \rfloor}\right) \tag{10}$$

This strategy aims to reduce exposure bias between training and inference conditions when predicted CUIs are imperfect. Finally, the `Salesforce/blip-image-captioning-large` [4] decoder is fine-tuned using prefix-based label masking, where prompt tokens are assigned the ignore index -100. Consequently, the autoregressive loss is computed only on generated caption tokens rather than on the conditioning prompt itself.

5. Experiments and Results

5.1. Dataset and Evaluation Metrics

All experiments were conducted using the an updated and extended version of the ROCOv2 dataset [3] provided by the ImageCLEFmedical Caption 2026 organizers [1]. Official benchmark evaluation is performed using a composite leaderboard score integrating both semantic relevance and factual consistency metrics.

The evaluation framework includes: *Overall*, *Relevance*, *Factuality*, *BERTScore* [5], *ROUGE* [10], *Image-Caption Similarity* [1], *BLEURT* [20], *MedCAT* [21], and *AlignScore* [11].

In addition to the official benchmark metrics, an internal clinical-oriented evaluation was conducted on the validation set using a custom dual-concept diagnostic dictionary consisting of *Anatomical Structures* and *Pathological Findings*. This additional evaluation is intended to better characterize clinically relevant generation behavior beyond conventional lexical similarity metrics.

For each category, the following metrics are reported:

- **Hallucination Rate:** proportion of images where at least one unsupported concept is generated;
- **Omission Rate:** proportion of images where at least one reference concept is missed;
- **Concept Precision**;
- **Concept Recall**;
- and **Concept F1-score**.

The dual-concept diagnostic dictionary was constructed by analyzing term frequency patterns across the ROCOv2 training captions. Anatomical concepts were organized into 25 semantic groups (e.g., *lung*, *heart*, *spine*, *blood_vessel*) covering the major anatomical regions represented in the dataset, with each group defined by a set of synonymous surface forms to account for terminological variation (e.g., *cardiac* and *myocardium* mapped to *heart*). Pathological findings were organized into three groups: *tumor_lesion* (mass, nodule, neoplasm), *fluid_effusion* (effusion, edema), and *lung_opacities* (opacity, consolidation, ground glass), reflecting the most frequently occurring pathological categories in the dataset.

Concept matching was performed using regular expression boundary matching (`\b`) after lowercasing, ensuring whole-word matches and avoiding partial string overlap. A concept was considered *hallucinated* if it appeared in the generated caption but not in the reference caption, and *omitted* if it appeared in the reference but not in the generated caption. Hallucination Rate and Omission Rate are reported at the image level, while Precision, Recall, and F1-score are aggregated at the concept level across the full validation set.

The pathological dictionary covers three semantic groups reflecting the most frequently occurring pathological categories in the ROCOv2 dataset. A more comprehensive coverage of pathological terminology, including findings such as fracture, stenosis, calcification, and infection, was not incorporated within the scope of the present study due to time constraints. Expanding the pathological dictionary to include a broader range of clinical findings represents a clear direction for future work, as a more exhaustive concept vocabulary would provide a more complete assessment of hallucination and omission behavior across the full spectrum of radiology pathologies.

5.2. Implementation Details

The proposed framework was implemented in PyTorch.

For the concept extraction module, the `EfficientNetCSRA` network was optimized using AdamW with an initial learning rate of 1×10^{-4} and a batch size of 32. Training was performed for 30 epochs using the two-phase optimization strategy described in Section 4, together with cosine annealing learning rate scheduling. The Asymmetric Loss [13] configuration used $\gamma_{\text{neg}} = 4.0$ and $\Delta = 0.05$.

For caption generation, input images were enhanced using CLAHE [24] and resized to 384×384 before being processed by the BLIP-large [4] backbone. The decoder was fine-tuned for 15 epochs using a learning rate of 2×10^{-5} together with the proposed dynamic prompt fading strategy.

5.3. Official Submission Results

Table 1 presents the official leaderboard performance of our submitted system (`DS-MED Lab`) on the ImageCLEFmedical Caption 2026 test set.

The proposed framework achieved an Overall score of 0.2814 and a Similarity score of 0.7286, reflecting strong visual-semantic alignment between the generated captions and the input images as measured by a CLIPScore-like formula using the MedImageInsights embedding model. This metric evaluates image-caption consistency rather than similarity to reference text, and therefore does not directly indicate lexical or semantic overlap with the ground-truth descriptions. However, the relatively lower factuality-oriented metrics highlight the continued difficulty of ensuring clinically precise generation in medical captioning systems.

These observations suggest that strong semantic similarity alone does not necessarily imply clinically reliable descriptions, motivating the need for complementary clinical-oriented evaluation protocols.

Table 1
ImageCLEFmedical Caption 2026 official leaderboard excerpt (selected systems by Overall score). The submitted system DS-MED Lab is highlighted.

Rank	Team	Overall	Relevance	Factuality	Bert	Rouge	Similarity	Bleurt	Medcat	Align
1	AlstatLab	0.3755	0.5786	0.1724	0.6250	0.3234	1.0104	0.3555	0.2160	0.1287
3	dsgt_caption	0.3571	0.5365	0.1777	0.6042	0.2809	0.9381	0.3229	0.1943	0.1610
5	csmorgan	0.3458	0.5377	0.1539	0.6075	0.2725	0.9408	0.3299	0.1751	0.1327
7	UIT_HighDefinition	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379
9	UMUTeam	0.2990	0.4798	0.1181	0.5670	0.2230	0.8147	0.3146	0.1048	0.1315
10	DS-MED Lab (Ours)	0.2814	0.4526	0.1102	0.5690	0.2126	0.7286	0.3004	0.1427	0.0778
11	krishnatewari	0.2685	0.4346	0.1025	0.5398	0.2059	0.6788	0.3137	0.1219	0.0831
Total participants: 13										

5.4. Internal Clinical Validation

To further analyze clinically relevant generation behavior, the proposed framework was evaluated using the dual-concept diagnostic dictionary described earlier. Table 2 compares the vanilla BLIP-large baseline against the proposed system.

Table 2
Internal clinical validation results. Hallucination Rate and Omission Rate are image-level metrics (lower is better). Precision, Recall, and F1-score are concept-level metrics (higher is better).

Model	Category	Halluc. Rate ↓	Omission Rate ↓	Precision ↑	Recall ↑	F1-Score ↑
Baseline BLIP-large	Anatomy	59.4%	49.0%	0.3789	0.4362	0.4056
	Pathology	32.2%	18.0%	0.2463	0.3583	0.2919
Ours (Full System)	Anatomy	**39.2%**	60.4%	**0.3808**	0.2437	0.2972
	Pathology	42.0%	**13.1%**	**0.2687**	**0.5312**	**0.3569**

The proposed framework substantially reduced anatomical hallucination from 59.4% to 39.2%, suggesting that the concept-guided prompting mechanism improved anatomical grounding during generation.

More importantly, pathology recall increased from 0.3583 to 0.5312, while pathology omission rate decreased from 18.0% to 13.1%. These results indicate that the proposed semantic guidance strategy improved the model's ability to surface clinically relevant pathological findings.

However, the improvements were accompanied by several trade-offs. In particular, pathology hallucination increased from 32.2% to 42.0%, and anatomy recall decreased from 0.4362 to 0.2437. This behavior suggests that strengthening sensitivity toward pathological findings may also increase the tendency to over-generate abnormal concepts under uncertain conditions.

Overall, the results highlight a broader tension between semantic completeness, hallucination suppression, and pathology sensitivity in medical image caption generation. While the proposed framework improves pathological coverage and reduces anatomical hallucination, further work is required to better balance sensitivity and specificity during clinically grounded generation.

latex

6. Discussion

The results presented in Tables 1 and 2 highlight an important limitation of current medical image captioning evaluation protocols. Although standard leaderboard metrics such as Similarity, BERTScore, ROUGE, and BLEURT are effective for measuring lexical and semantic similarity, they do not explicitly penalize clinically incorrect statements. Consequently, a caption may achieve strong overlap with reference text while still hallucinating anatomical structures or omitting clinically relevant findings.

The experimental results suggest that optimizing caption generation with explicit clinical priors can partially mitigate this issue. By integrating concept-aware routing and CUI-guided prompting, the proposed framework substantially improves pathological coverage, increasing Pathology Recall from 0.3583 to 0.5312 while reducing anatomical hallucination from 59.4% to 39.2%. These improvements

indicate that the model becomes more conservative in generating unsupported anatomical content and more sensitive to clinically meaningful abnormalities.

However, this improvement is accompanied by a decrease in anatomical recall, reflecting a trade-off between exhaustive anatomical description and pathology-focused reporting. This trade-off is considered clinically reasonable in the context of radiology assistance systems, where missing a true lesion is generally more critical than omitting redundant anatomical context. The results therefore expose a broader mismatch between conventional language-generation objectives and clinically reliable reporting.

Relationship Between Clinical and Official Evaluation. A notable observation is that improvements in the clinical-oriented evaluation do not translate directly into gains on the official factuality metrics. Despite reducing anatomical hallucination and improving pathology recall, the AlignScore and MedCAT scores remain low (0.0778 and 0.1427 respectively). This discrepancy arises because AlignScore measures textual entailment consistency while MedCAT evaluates UMLS concept overlap at a finer granularity than the curated dictionary. The gap between these two evaluation perspectives highlights the need for unified clinical evaluation frameworks that jointly assess concept coverage, factual entailment, and pathology-specific recall.

Handling Incompatible Domain-Concept Predictions. Since macro-domain routing and CUI prediction are performed independently, the framework may occasionally produce semantically inconsistent prompt combinations — for example, routing an image to the cardiology macro-domain while simultaneously predicting bone fracture concepts. In such cases, the textual prompt presented to the BLIP-large decoder contains contradictory semantic signals.

It is reasonable to expect that the autoregressive decoder partially absorbs these inconsistencies: rather than faithfully reproducing all conflicting concepts, the decoder may tend to prioritize tokens that align with the visual evidence encountered during fine-tuning, though a systematic analysis of this behavior remains an avenue for future work. Nevertheless, this implicit error correction role places an additional burden on the decoder beyond pure text generation, and may partly explain the residual factual errors observed in the official evaluation, particularly the low AlignScore and MedCAT scores reported in Table 1.

More broadly, when the concept extractor produces incorrect CUIs, the autoregressive decoder is conditioned on flawed semantic priors, which can propagate factual errors into the generated caption regardless of decoder quality. This propagation mechanism represents a fundamental limitation of pipeline-based concept-guided generation and motivates future work on joint training or uncertainty-aware prompt filtering to reduce the impact of upstream concept errors on downstream caption fidelity.

The findings demonstrate that concept-guided prompting and clinical-aware routing provide a practical direction for reducing hallucinations in medical vision-language models. While the proposed system does not fully eliminate factual errors, it establishes a stronger alignment between generated captions and clinically relevant image content, which is essential for trustworthy medical AI systems.

7. Future Work

Several directions may further improve the clinical reliability and robustness of medical image captioning systems. First, future research should investigate evaluation protocols that better capture factual correctness and diagnostic relevance rather than relying primarily on lexical overlap metrics. Integrating clinically grounded metrics such as RadGraph F1 [25], CheXpert [26] label agreement, or entity-level factual consistency into both evaluation and training objectives may provide stronger alignment with real-world radiological reporting requirements.

Second, the current framework relies on a constrained subset of UMLS concepts for visual grounding. Expanding the concept vocabulary to include finer-grained pathological findings, anatomical relations, and procedural terminology could improve the expressiveness and diagnostic completeness of generated

captions. More advanced concept linking and ontology-aware reasoning mechanisms may also help reduce semantic ambiguity during generation.

Third, although our cluster-aware prompting strategy improves pathology recall, the system still exhibits residual hallucination and omission errors. Future work could explore retrieval-augmented generation, external medical knowledge integration, or uncertainty-aware decoding strategies to further enhance factual reliability. In particular, incorporating radiology-specific structured representations may help the model maintain stronger consistency between visual evidence and generated text.

Finally, recent large-scale medical vision-language models such as LLaVA-Med [27], BioViL-T [28], and instruction-tuned multimodal LLMs provide promising foundations for clinically grounded caption generation. Investigating how explicit concept routing and hallucination-aware prompting can be combined with these larger architectures represents an important direction for developing more trustworthy medical AI systems.

Acknowledgments

This research was funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number CS4-2026-01.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and Gemini (Google) in order to: grammar and spelling check, paraphrase and reword, and language refinement of manuscript drafts. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Damm, H., Pakull, T. M. G., Reinartz, L., Bracke, B., Eryılmaz, B., Nath, P., Brüngel, R., Schmidt, C. S., Schäfer, H., Ben Abacha, A., García Seco de Herrera, A., Müller, H., and Friedrich, C. M. Overview of ImageCLEFmedical 2026 – Medical Concept Detection and Caption Generation with Synthetic Data Extensions. In *CLEF 2026 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 21–24, 2026. (To appear)

[2] Ionescu, B., Müller, H., Stanciu, D.-C., Radu, A., Bolborici, R.-G., Negru, M., Ene, A.-F., Vasilescu, V.-M., Nicolae, A.-A., Ştefan, L.-D., Constantin, M.-G., Dogariu, M., Andrei, A.-G., Damm, H., Pakull, T. M. G., Ben Abacha, A., García Seco de Herrera, A., Friedrich, C. M., Brüngel, R., Reinartz, L., Schäfer, H., Schmidt, C. S., Bracke, B., Nath, P., Eryılmaz, B., Hjuler, M., Fabre, D., Lemaire, C., Lecouteux, B., Schwab, D., Dimitrov, D., Hee, M. S., Ahsan, M., Ahmad, S., Zlatkova, D., Pachov, G., Xie, Z., Nakov, P., Koychev, I., Heras Rivera, J. E., Low, D. K., Yim, W., Ruzevick, J., Child, D., Kurt, M., Sun, Z., Xia, F., Yetisgen, M., Radzhabov, A., Prokopchuk, Y., Kovalev, V., Karpenka, D., Hicks, S. A., Gautam, S., Riegler, M. A., Thambawita, V., Halvorsen, P., El Sakka, M., Mothe, J., Băicoianu, A., Florea, C.-N., and Ivanovici, M. Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction – Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science, Jena, Germany, September 21–24, 2026.

[3] Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., Koitka, S., Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Müller, H., Horn, P., Nensa, F., and Friedrich, C. M. ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset. *Scientific Data*, 11(1), 2024. https://doi.org/10.1038/s41597-024-03496-6

[4] Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[5] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

[6] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

[7] Luong, M. V., Dinh-Doan, M. H., Bui-Hoang, G. P., and Nguyen-Tat, T. B. UIT-Oggy at ImageCLEFmedical 2024 Caption: CSRA-Enhanced Concept Detection and BLIP-Driven Vision-Language Captioning. In *CLEF 2025 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[8] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. *NEJM AI*, 2(1), 2024. https://doi.org/10.1056/AIoa2400640

[9] Tan, M. and Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, pp. 6105–6114, Long Beach, CA, 2019.

[10] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, Barcelona, Spain, 2004.

[11] Zha, Y., Yang, Y., Li, R., and Hu, Z. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 11328–11348, 2023.

[12] Zhu, K. and Wu, J. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 184–193, 2021.

[13] Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric Loss for Multi-Label Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, 2021.

[14] Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Bracke, B., Damm, H., Pakull, T. M. G., Schmidt, C. S., Müller, H., and Friedrich, C. M. Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection. In *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, pp. 1437–1455, 2024.

[15] Damm, H., Pakull, T. M. G., Becker, H., Bracke, B., Eryılmaz, B., Bloch, L., Brüngel, R., Schmidt, C. S., Rückert, J., Pelka, O., Schäfer, H., Idrissi-Yaghir, A., Ben Abacha, A., García Seco de Herrera, A., Müller, H., and Friedrich, C. M. Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation. In *CLEF 2025 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, pp. 2197–2223, 2025.

[16] Phan, T. V., Nguyen, T. K., Hoang, Q. A. D. D., Phan, Q. T., and Nguyen-Tat, T. B. UIT-2Q2T at ImageCLEFmedical 2024 Caption: Multimodal Medical Image Captioning using Bootstrapping Language-Image Pre-training. In *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, pp. 1711–1719, 2024.

[17] Nguyen, Q. V., Pham, H. Q., Tran, D. Q., Nguyen, T. K.-B., Nguyen-Dang, N.-H., and Nguyen-Tat, T. B. UIT-DarkCow Team at ImageCLEFmedical Caption 2024: Diagnostic Captioning for Radiology Images Efficiency with Transformer Models. In *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, pp. 1695–1710, 2024.

[18] Pan, R., Bernal Beltrán, T., García Díaz, J. A., and Valencia-García, R. UMUTeam at ImageCLEF 2025: Fine-Tuning a Vision-Language Model for Medical Image Captioning and SapBERT-Based Reranking for Concept Detection. In *CLEF 2025 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[19] Nguyen, N. N.-Y., Le Tu, H., Nguyen, P. D., Do, T. N., Thai, T. M., and Nguyen-Tat, T. B. DS@BioMed at ImageCLEFmedical Caption 2024: Enhanced Attention Mechanisms in Medical Caption Gen-

eration through Concept Detection Integration. In *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, pp. 1681–1694, 2024.

[20] Sellam, T., Das, D., and Parikh, A. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7881–7892, 2020.

[21] Kraljevic, Z., Searle, T., Shek, A., et al. Multi-domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083, 2021.

[22] Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.

[23] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297, 1967.

[24] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. Adaptive Histogram Equalization and Its Variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[25] Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D., Bui, T., Chambon, P., Zhang, Y., Lungren, M., Ng, A., Langlotz, C., and Rajpurkar, P. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1, 2021. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf

[26] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M., and Ng, A. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 590–597, 2019. https://doi.org/10.1609/aaai.v33i01.3301590

[27] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 28541–28564, 2023. https://doi.org/10.5555/3666122.3667550

[28] Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M., Nori, A., Alvarez-Valle, J., and Oktay, O. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19293–19304, 2023. https://doi.org/10.1109/CVPR52729.2023.01850

A. Online Resources

To ensure full reproducibility of our results, the source code for models developed in this study is publicly available on GitHub. The repositories are organized as follows: https://github.com/nguyenthinh1406/ImageCLEF2026_image_captioning.