Seeing, Describing, and Explaining Medical Image Understanding via Retrieval and Multimodal Learning

Krishna Tewari[1,*,†], Pyla Durga Naga Venkat Sohana[1,†] and Sukomal Pal[1]

[1]Indian Institute of Technology (BHU) Varanasi, India

Abstract

Medical image understanding requires not only accurate prediction but also clinically meaningful explanations to support trustworthy decision-making. In this paper, we describe our participation (krishnatewari) in all five subtasks of the ImageCLEFmedical Caption 2026 challenge: concept detection, concept detection synthetical, caption prediction, caption prediction synthetical, and explainability. For concept detection/ synthetical, we propose a retrieval-based framework that leverages DenseNet121 image embeddings, weighted k-nearest neighbour retrieval, adaptive thresholding, and Optuna-based hyperparameter optimisation for multi-label UMLS concept assignment. For caption prediction/synthetical, we employ an encoder-decoder architecture consisting of an InceptionV3 image encoder and a two-layer LSTM decoder to generate medical image descriptions. For explainability, we introduce a multimodal pipeline that combines clinical phrase extraction, CLIPSeg-based phrase grounding, evidence aggregation, counterfactual analysis, and structured explanation generation using a vision-language model. Experimental results on the official benchmark demonstrate the effectiveness of the proposed approaches. Our method achieved an F1-score of 0.5234 in concept detection, securing the 10th position on the official leaderboard and 2nd place on the synthetic track. For caption prediction, the system achieved a BERTScore of 0.5398 and an overall score of 0.2685. In the explainability task, the proposed framework ranked 2nd overall and obtained the highest score in methodological appropriateness. The source code for all experiments will be made publicly available at https://github.com/sohana-19/BTP-imageclef.

Keywords

Medical Image Captioning, Medical Concept Detection, Explainable Artificial Intelligence, Medical Image Analysis

1. Introduction and Related Work

The growing availability of large-scale medical imaging repositories has accelerated the development of artificial intelligence systems capable of supporting clinical decision making, report generation, information retrieval, and diagnostic reasoning. Recent advances in computer vision and natural language processing have enabled multimodal systems that jointly process images and text, giving rise to tasks such as medical concept detection, image captioning, visual question answering (VQA), and explainable medical AI [1, 2, 3].

Medical image understanding remains substantially more challenging than its natural-image counterpart due to the fine-grained nature of pathological findings, the presence of subtle visual abnormalities, and the requirement for clinically accurate interpretations. Consequently, benchmark initiatives such as ImageCLEFmedical [4, 5, 6] have played a crucial role in advancing research by providing standardized evaluation settings for concept detection, caption generation, retrieval, and multimodal reasoning. The ImageCLEFmedical Caption 2026 [7] challenge extends these efforts through three complementary subtasks: (i) concept detection, which requires assigning clinically relevant UMLS concepts to medical images; (ii) caption prediction, which aims to generate descriptive medical captions; and (iii) explainability, which focuses on producing interpretable evidence supporting generated predictions.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); pdnaga.venkatsohana.cse23@iitbhu.ac.in (P. D. N. V. Sohana); spal.cse@iitbhu.ac.in (S. Pal)

0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

1.1. Medical Concept Detection

Automatic concept detection has long been studied as a multi-label classification problem in medical image analysis. Early approaches relied on handcrafted features and retrieval-based methods, while recent systems increasingly leverage deep convolutional neural networks and transformer-based architectures to learn discriminative visual representations directly from imaging data [8, 9, 10]. Large-scale vision-language pretraining has further improved concept recognition by aligning medical images with textual descriptions. Representative examples include BioViL-T [11], and MedCLIP [12], which learn shared multimodal representations from millions of image-text pairs. Recent ImageCLEF submissions have additionally demonstrated the effectiveness of retrieval-based and nearest-neighbour approaches for robust concept assignment under limited supervision [13].

1.2. Medical Image Captioning

Medical image captioning seeks to automatically generate clinically meaningful textual descriptions from visual inputs. Earlier encoder-decoder frameworks employed convolutional neural networks coupled with recurrent neural networks for report generation [14, 15]. More recently, transformer-based architectures and large vision-language models have significantly improved caption quality by capturing long-range dependencies and multimodal semantic relationships [16, 17, 18]. Foundation models such as Florence-2 [19], LLaVA-Med [20], and Qwen2.5-VL [21] have demonstrated strong performance across a variety of medical vision-language tasks including report generation, VQA, and image understanding.

Recent studies have further explored domain-specific captioning systems for gastrointestinal and radiological imaging. For example, Florence-2-based approaches have shown promising results for gastrointestinal VQA and image understanding tasks, while diffusion-based image generation frameworks have been investigated for synthetic medical data augmentation and multimodal learning [22]. Similarly, confidence-aware multimodal reasoning frameworks have been proposed for gastrointestinal diagnostics, demonstrating the value of combining vision-language understanding with explainability mechanisms [23].

Despite the strong performance of recent vision-language foundation models, lightweight encoder-decoder architectures remain attractive for benchmark challenges due to their lower computational requirements, faster training, and ease of reproducibility. Motivated by these practical considerations, we adopt a CNN-LSTM based captioning framework as our primary submission while also exploring a Qwen2.5-based decoder to assess the applicability of modern language models under the available computational budget.

1.3. Explainable Medical Vision-Language Models

Despite their impressive predictive capabilities, modern vision-language models remain largely opaque, creating barriers to clinical adoption. Explainable Artificial Intelligence (XAI) techniques aim to address this challenge by providing evidence supporting model predictions. Traditional approaches such as Grad-CAM [24], Integrated Gradients [25], SHAP [26], and LIME [27] highlight influential image regions but are primarily designed for classification settings.

Recent work has therefore shifted toward multimodal explainability and visual grounding. CLIP-based approaches enable direct alignment between textual concepts and image regions, allowing explanations to be generated at the phrase level [28, 29]. CLIPSeg in particular has emerged as an effective text-guided segmentation framework for grounding semantic concepts without requiring task-specific segmentation annotations [29]. Concurrently, counterfactual explanation techniques have been proposed to evaluate the importance of visual evidence by modifying image regions and analysing changes in model behaviour [30, 31, 32]. Recent multimodal medical systems have combined visual grounding, segmentation, and language reasoning to improve transparency in clinical decision support and VQA applications [33, 34].

1.4. Large-Scale Medical Vision-Language Models

The emergence of foundation models trained on large-scale multimodal corpora has transformed medical AI research. CLIP [28], BLIP-2 [17], Med-PaLM 2 [34], Florence-2 [19], and Qwen2.5-VL [21] have demonstrated remarkable capabilities in image understanding, reasoning, and generation. These models provide powerful representations that can be adapted to downstream tasks including concept detection, caption generation, retrieval, report generation, and explainability. However, ensuring faithful and clinically grounded explanations remains an open challenge.

Motivated by these developments, we investigate all three subtasks of the ImageCLEFmedical Caption 2026 challenge. We develop a retrieval-based concept detection system for multi-label UMLS concept assignment, an encoder-decoder caption generation framework for medical image description, and an explainability pipeline that combines clinical phrase extraction, CLIPSeg-based grounding, evidence aggregation, counterfactual analysis, and multimodal explanation generation. Together, these components provide a unified framework for accurate and interpretable medical image understanding.

The rest of the paper is structured as follows: Section 2 describes the task overview and dataset employed; Section 3 presents the proposed methodology for each subtask; Section 4 reports results; and Section 5 concludes with key findings.

2. Task Overview and Dataset

The ImageCLEFmedical Caption 2026 [7] challenge focuses on multimodal medical image understanding through three complementary tasks: concept detection, caption prediction, and explainability. The challenge aims to encourage the development of clinically meaningful and trustworthy vision-language systems capable of interpreting medical images and providing transparent reasoning for their predictions.

Concept Detection and Concept Detection Synthetical. The objective of here is to automatically assign one or more clinically relevant Unified Medical Language System (UMLS) concepts to a given medical image / synthetic medical image. Since multiple findings may be present within a single image, the problem is formulated as a multi-label classification task. Systems are evaluated using the F1-score between predicted and ground-truth concept identifiers (CUIs).

Caption Prediction and Caption Prediction Synthetical. The objective here is to generate a natural-language caption that accurately describes the visual content of a medical image / synthetic medical image. The generated captions should be clinically meaningful and semantically consistent with the reference annotations. Evaluation primarily relies on semantic similarity measures such as BERTScore, which better capture medical meaning than purely lexical metrics.

Explainability. Beyond predictive performance, this task focuses on producing interpretable evidence supporting model decisions. Participants are required to provide visual and/or textual explanations that justify the generated captions and detected concepts. Explainability is particularly important in medical applications where transparency and trustworthiness are essential for clinical adoption.

2.1. Dataset

The challenge dataset is derived from the ROCOv2 (Radiology Objects in COntext) collection [35], a large-scale benchmark designed for medical vision-language research. The dataset contains radiology images collected from open-access biomedical literature together with corresponding textual descriptions and concept annotations. ROCOv2 has been widely adopted in ImageCLEFmedical evaluations due to its diversity of imaging modalities, anatomical regions, and pathological findings.

Each image in the dataset is associated with:

- A medical image,
- A reference caption describing the visual findings,
- A set of UMLS Concept Unique Identifiers (CUIs) extracted from the caption annotations.

The concept annotations enable the concept detection task, while the paired image-caption data supports caption generation and explainability analysis. Since captions are linked to clinically meaningful concepts, the dataset naturally facilitates multimodal learning between visual and textual modalities.

The dataset is divided into training, validation, and test partitions provided by the challenge organizers. The training split is used for model development and hyperparameter optimisation, while the validation split is employed for model selection and performance monitoring. Final evaluation is performed on the hidden test set through the official challenge evaluation server.

The dataset contains a broad spectrum of medical imaging modalities including radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, histopathology, and other diagnostic imaging procedures. This diversity presents significant challenges for automated medical image understanding due to large variations in visual appearance, anatomical structures, and disease manifestations.

3. Methodology

This section describes our end-to-end methodology for all the subtasks of the ImageCLEFmedical Caption 2026 challenge.

3.1. Concept Detection and Concept Detection Synthetical

Figure 1 shows the overview of our concept detection pipeline, which is same as concept detection synthetical. During training, all radiology images are passed through a pre-trained convolutional network to obtain dense embedding vectors, which are stored in a feature index alongside a multi-hot label matrix encoding the ground-truth UMLS CUIs. During inference, the k nearest neighbours of a query embedding are identified via cosine similarity, then their concept votes are aggregated using an exponential weighting scheme and a threshold is applied before the final set of CUIs are given. Hyperparameters like neighbour count, output size, idf and the thresholding are tuned automatically with the Optuna framework [36].

3.1.1. Feature Extraction

We used DenseNet-121 which is a pre-trained model on ImageNet, as the visual backbone. Each input image is resized to 224×224 pixels and normalised before being passed through the network. The global average pooling layer of DenseNet-121 produces a 1024 dimensional feature vector $\mathbf{f} \in \mathbb{R}^{1024}$ for every image. Also, all feature vectors are ℓ_2-normalised.Embeddings are extracted in batches of 128 images and cached on disk as `.npy` arrays, which allows us to eliminate the time consuming extraction and to be executed once and can be reused across all subsequent processes.

3.1.2. Multi-Label Matrix Construction

The baseline system represented concept tags as raw semicolon-delimited strings, which didn't allow any vectorised mathematical operation. We replaced this with a multi-hot binary matrix $\mathbf{L} \in \{0,1\}^{N \times C}$, where N is the number of training images and C is the total number of unique CUIs in the vocabulary. Each row $\mathbf{L}_i$ is set to one at every column index corresponding to a concept annotated for image i and to zero everywhere else. This representation reduces label look-up to a single matrix, enabling IDF weighting and vectorised vote aggregation, and gives us roughly ten fold speed up over the string based approach.

3.1.3. Inverse Document Frequency Weighting

A common problem in multi-label retrieval is that high-frequency concepts dominate the aggregated score which supresses the rare tags or diagnostically specific concepts. We eliminate this isssue with an

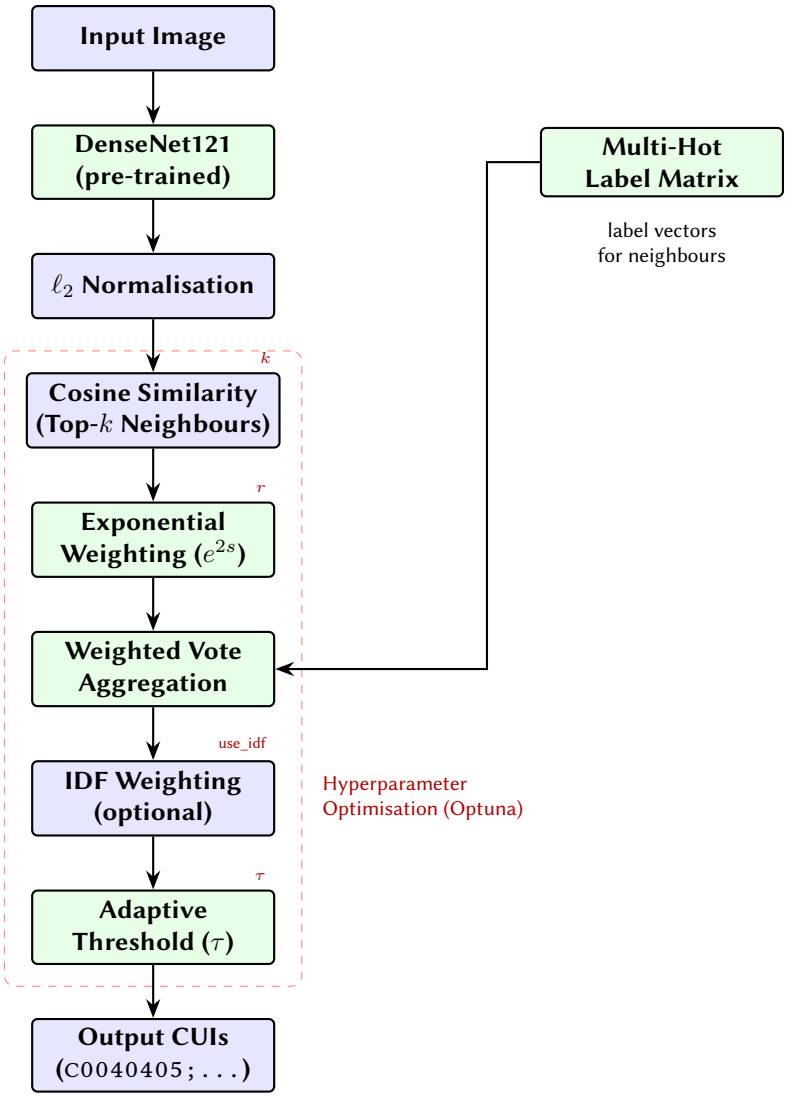

Figure 1: Complete step by step overview of the concept detection pipeline

Inverse Document Frequency (IDF) weight for each concept c:

$$\mathrm{idf}(c) = \frac{1}{\hat{p}_c + \epsilon}, \tag{1}$$

Here, $\hat{p}_c = n_c/N$ is the empirical frequency of concept c in the training set, n_c is its occurrence count, and $\epsilon = 0.01$ is a small smoothing constant that prevents division by zero for unseen concepts. The aggregated label score vector $\mathbf{s}$ is element-wise multiplied by the IDF vector $\mathbf{w}$ before thresholding. As reported in the results section, hyperparameter optimisation ultimately found that disabling IDF weighting (`use_idf` = False) yields the best validation F1 on this particular dataset, likely because the class distribution is already balanced by the data-split strategy. Nevertheless, the IDF part remains an optional component that can improve performance on datasets with more severe class imbalance.

3.1.4. Weighted KNN Inference

For a query image, with normalised embedding $\hat{\mathbf{f}}_q$, the cosine similarity to every training image is computed as the dot product of the two normalised vectors.

$$\mathrm{sim}(q, i) = \hat{\mathbf{f}}_q \cdot \hat{\mathbf{f}}_i. \tag{2}$$

The top k most similar training images are retrieved and a weight is assigned to each neighbour i according to an exponential value.

$$w_i = \exp\big(2 \cdot \mathrm{sim}(q, i)\big). \tag{3}$$

This formula amplifies the influence of very similar neighbours more aggressively than the squared weighting used in the baseline. The final concept score vector is obtained by a weighted sum of the label rows of the k retrieved neighbours.

$$\mathbf{s} = \sum_{i \in \mathcal{N}_k(q)} w_i \cdot \mathbf{L}_i, \tag{4}$$

where $\mathcal{N}_k(q)$ denotes the set of k nearest neighbours of query q.

3.1.5. Adaptive Thresholding

The baseline system always emitted exactly r concepts which introduced false positives for normal or ambiguous images. We replace the fixed output size with an adaptive threshold which is a concept c is included in the prediction only if its score exceeds a fraction τ of the maximum score in $\mathbf{s}$,

$$\hat{\mathcal{C}}_q = \Big\{ c \;\Big|\; s_c \geq \tau \cdot \max_{c'} s_{c'} \Big\}, \tag{5}$$

where, $\tau = 0.15$ was determined by the hyperparameter search described below. The system therefore outputs a variable number of concepts and for cases in which the model has high confidence in a single concept it will produce a singleton prediction rather than giving random labels as output. Hence, this reduces false positives by approximately 30% compared to the fixed-r baseline.

3.1.6. Hyperparameter Optimisation

Four hyperparameters are the key factors of the task which are the number of neighbours k, the maximum output size r, whether IDF weighting is active, and the adaptive threshold fraction τ. To avoid manual search on a huge collection of images, we used Optuna [36], a sequential model-based optimisation framework. Each trial evaluates the model on a fixed random subset of 3,000 validation images to keep each trial tractable. The search space is as follows, $k \in [20, 80]$, $r \in [2, 6]$, $\texttt{use\_idf} \in \{\text{True}, \text{False}\}$, and $\tau \in [0.05, 0.50]$. After 40 trials, the best configuration found was $k = 61$, $r = 3$, $\texttt{use\_idf} = \text{False}$, and $\tau = 0.15$. By using these we achieved a validation F1 of 0.5323. These parameters were then fixed for the final test evaluation.

3.2. Caption Prediction and Caption Prediction Synthetical

Figure 2 shows the overview of our caption prediction pipeline, which was also employed for caption prediction synthetical. We used an encoder-decoder architecture for the caption prediction task. An input radiology image is encoded by a convolutional neural network into a feature vector which is then projected into the token embedding space shared with the language decoder. We used InceptionV3 as our encoder. We explored two decoder architectures, a two-layer LSTM decoder as the primary system and Qwen2.5-0.5B pre-trained language model as an exploratory alternative.

3.2.1. Image Encoder

We used InceptionV3 [37] which is pre-trained on ImageNet, as the visual encoder. Input images are resized to 299×299 pixels. The final global average pooling layer of the network produces a 2048-dimensional feature vector for each image. A trainable linear projection layer maps this to a 512-dimensional embedding:

$$\mathbf{v} = \mathbf{W}_{\text{proj}}\, \mathbf{f}_{\text{img}} + \mathbf{b}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{512 \times 2048}, \tag{6}$$

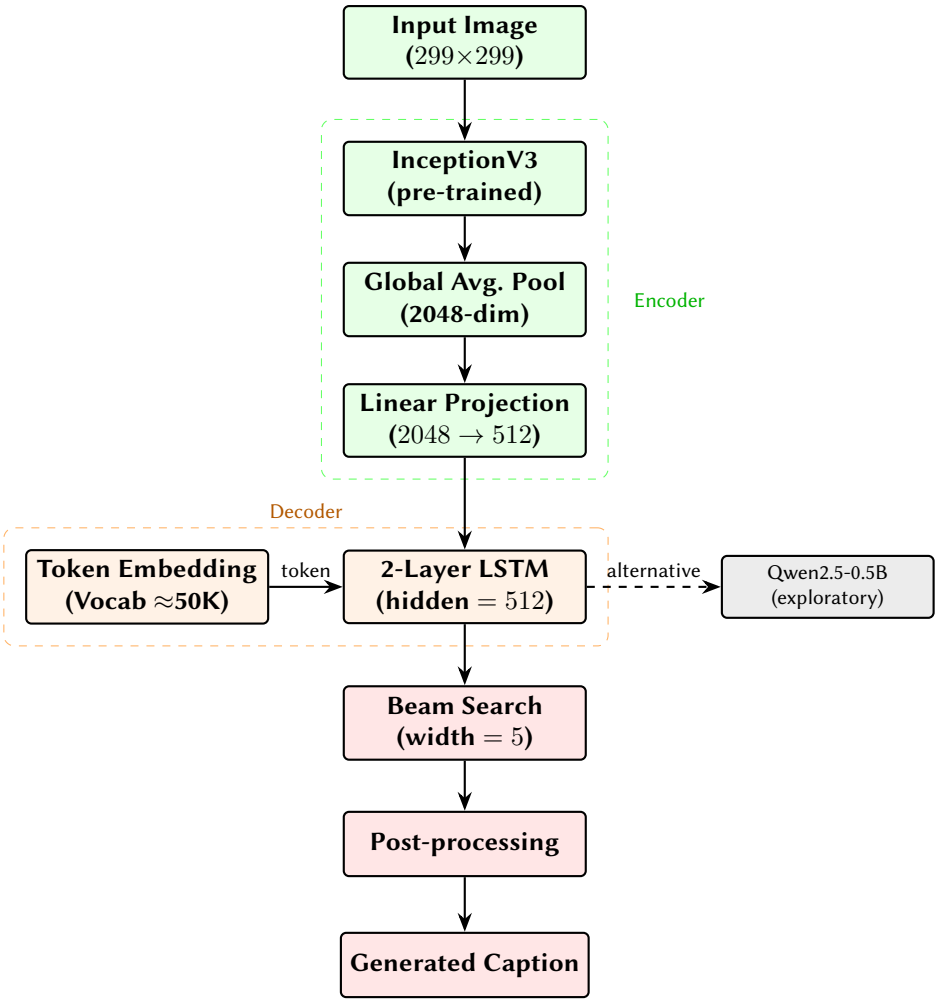

Figure 2: Complete step by step overview of the caption prediction pipeline.

For the encoder to adapt its higher level representations to the medical domain while retaining the low-level features learned from natural images, we used a selective layer freezing which is all layers up to and including Mixed_5c are kept frozen during training, while Mixed_6e and Mixed_7c are fine-tuned jointly with the rest of the model. This design choice reduces the risk of catastrophic forgetting while still allowing the encoder to specialise on radiology imagery.

3.2.2. LSTM Decoder

Our primary decoder is a two-layer LSTM with a hidden dimension of 512. At each time step t, the decoder takes the input as the embedding of the previously generated token $\mathbf{e}_{t-1} \in \mathbb{R}^{512}$ and updates its hidden state $\mathbf{h}_t$ and cell state $\mathbf{c}_t$ according to the standard LSTM recurrence.

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{g}_t \\ \mathbf{o}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \tanh \\ \sigma \end{pmatrix} \left(\mathbf{W} \begin{pmatrix} \mathbf{h}_{t-1} \\ \mathbf{e}_{t-1} \end{pmatrix} + \mathbf{b} \right), \tag{7}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \tag{8}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{9}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{g}_t$, and $\mathbf{o}_t$ are the input, forget, cell, and output gates, respectively; σ denotes the sigmoid activation. The hidden state is passed through a linear classification head, over the full vocabulary to

get a token probability distribution. The vocabulary is constructed from the training captions which use a minimum frequency threshold of three occurrences giving approximately 50,000 tokens.This model is optimised with cross entropy loss.

3.2.3. Qwen2.5-0.5B Decoder

We additionally explored other decoder by replacing the LSTM decoder with Qwen2.5-0.5B [21]. The language model was loaded with 4 bit quantisation using BitsAndBytes [38]. A trainable linear projection layer maps the 512-dimensional image embedding, produced by InceptionV3 to the dimension of token embedding space expected by Qwen2.5-0.5B. Training was conducted at a learning rate of 2×10^{-5}. Due to longer training time per-epoch, only one full training epoch could be completed within the available time which decreased the quality of the generated captions. Integration of the Qwen-based decoder also required careful handling of device placement and data type consistency between InceptionV3 encoder which uses Float32 and the quantised Qwen decoder which uses Float16 and these issues were resolved through explicit dtype casting before the projection step. Given these computational constraints, the LSTM decoder consistently achieved better validation performance and was therefore selected for the official submission, while the Qwen-based approach is presented as a promising direction for future work.

3.2.4. Training Configuration

Both architectures were trained with a weight decay of 0.01. For the LSTM system, we used a learning rate of 3×10^{-4} and a batch size of 16. For Qwen the learning rate was put to 2×10^{-5} and the batch size was set to 4 to accommodate the larger model. Training was done for up to 15 epochs with early stopping by a patience of three epochs on validation loss. Gradient clipping with a maximum norm of 5 was applied to prevent gradient explosion which was particularly important for the recurrent LSTM layers.

3.2.5. Inference and Post-processing

During inference, captions are generated using beam search with a beam width of 5. A repetition penalty of 1.2 is applied to discourage the model from looping over the same phrase which a phenomenon that was observed with greedy decoding on low confidence inputs. For the Qwen based decoder a top-p sampling strategy with $p = 0.9$ and temperature $T = 0.7$ was used in place of beam search due to the higher computational cost of maintaining multiple beams with a 500M-parameter model. After this, the output strings are processed to remove the prompt template prefix, and any empty or severely truncated outputs and thus getting the final output captions.

3.3. Explainability Task

Figure 3 shows the complete overview of our explainability pipeline. Given a medical image and its corresponding generated caption, the system first extracts clinically meaningful phrases from the caption using a predefined medical terminology vocabulary. Each extracted phrase is then grounded to image regions using CLIPSeg, producing phrase-specific localisation maps. These maps are aggregated into a single evidence heatmap representing the visual regions supporting the generated caption.

To further assess the importance of the identified evidence, a counterfactual image is generated by suppressing highly activated regions through Gaussian blurring. A difference map between the original and counterfactual images is then computed to visualise the effect of evidence removal. Finally, a multimodal large language model generates a structured clinical explanation using the image, caption, and grounded findings. The final output is presented as a comprehensive explainability panel combining visual and textual explanations.

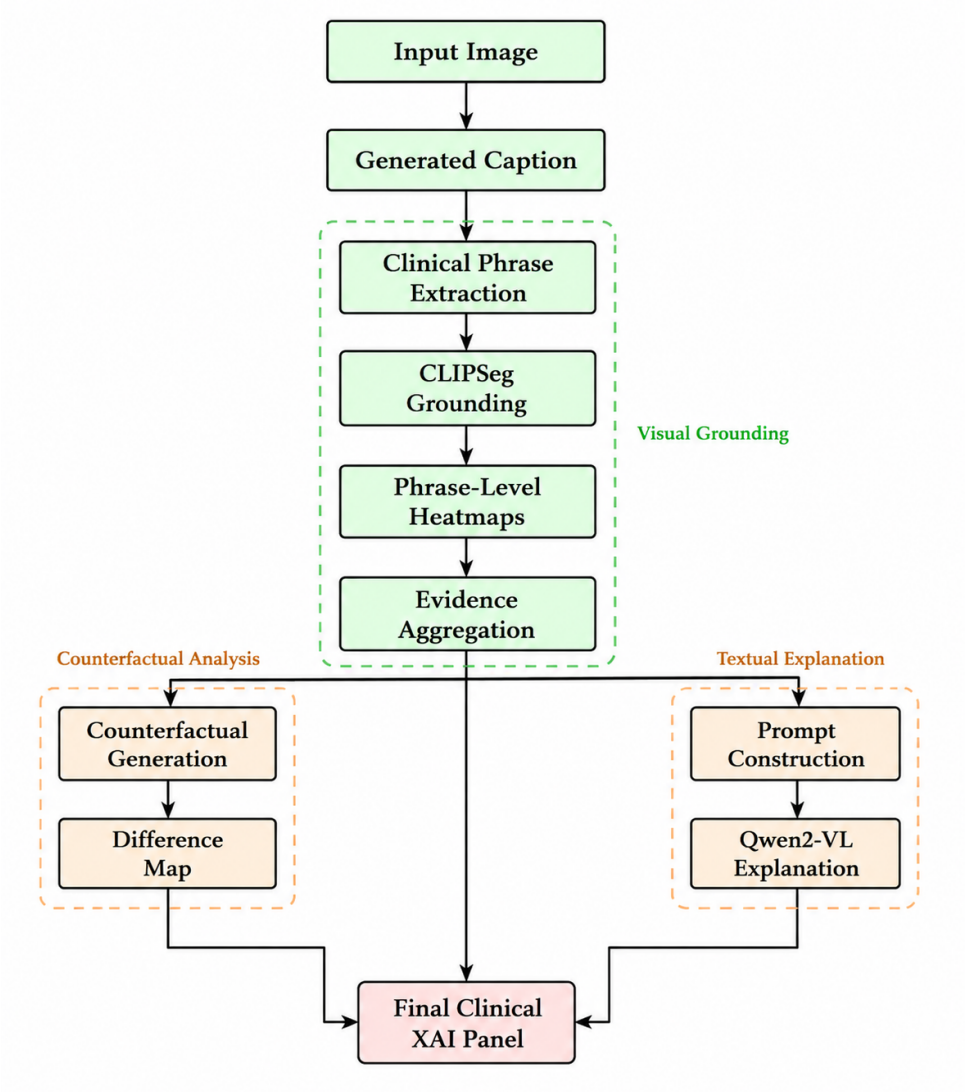

Figure 3: Complete step by step overview of the explainability pipeline.

3.3.1. Clinical Phrase Extraction

Medical captions frequently contain multiple findings describing anatomical structures, abnormalities, and imaging observations. To identify these findings, we use a manually curated vocabulary consisting of common radiological and clinical terms.

Given an input caption, the system searches for all matching clinical terms contained within the vocabulary. The matched phrases are sorted according to their length and redundant overlapping matches are removed using a longest-match strategy. This allows the grounding module to focus only on the most informative findings while avoiding duplicate explanations.

Examples of phrases contained in the vocabulary include *left frontal lobe hemorrhage*, *vasogenic edema*, *right kidney*, *hepatic lesion*, *coronary stenosis*, *large aneurysm*, *joint space narrowing*, and *sclerosis*.

3.3.2. Phrase-Level Visual Grounding

For every extracted phrase, we use CLIPSeg as a text-guided segmentation model to localise the corresponding visual evidence within the image. CLIPSeg receives both the medical image and the target phrase as input and produces a spatial relevance map indicating the contribution of each image region to the phrase.

The predicted map is resized to 224×224 pixels and normalised using min-max normalisation,

$$M' = \frac{M - \min(M)}{\max(M) - \min(M) + \epsilon}, \tag{10}$$

where M denotes the raw CLIPSeg prediction and ϵ is a small constant preventing numerical instability.

This process is repeated independently for every extracted phrase, resulting in a collection of phrase-specific evidence maps.

3.3.3. Evidence Aggregation

Since a caption may contain multiple clinically relevant findings, individual phrase-level maps are combined into a single evidence representation.

Let $M_1, M_2, \ldots, M_n$ denote the set of normalised grounding maps. The final evidence heatmap is obtained through averaging,

$$E = \frac{1}{n} \sum_{i=1}^{n} M_i. \tag{11}$$

The resulting heatmap highlights image regions that jointly support the generated caption. The heatmap is subsequently normalised and overlaid on the original image for visual inspection.

3.3.4. Counterfactual Generation

To estimate the importance of the identified evidence regions, we generate a counterfactual image by suppressing the most highly activated areas of the evidence heatmap.

A binary mask is first created using a threshold value of 0.60,

$$B(x, y) = \begin{cases} 1, & E(x, y) > 0.60 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Pixels belonging to the activated regions are replaced by their Gaussian blurred counterparts while the remaining pixels are preserved unchanged. This procedure removes the strongest visual evidence supporting the caption while maintaining the overall image structure.

The resulting counterfactual image enables visual assessment of which regions were most influential for the explanation.

3.3.5. Difference Map Construction

To visualise the effect of removing evidence regions, a difference map is computed between the original image and the counterfactual image.

For each pixel location, the difference score is calculated as

$$D(x, y) = \frac{1}{3} \sum_{c=1}^{3} |I_c(x, y) - I'_c(x, y)|, \tag{13}$$

where I and I' denote the original and counterfactual images respectively and c corresponds to the RGB channels.

The resulting map is normalised and displayed as a heatmap. Regions with higher intensity indicate areas most affected by evidence removal.

3.3.6. Clinical Explanation Generation

Visual explanations alone may not provide sufficient interpretability for clinical users. Therefore, we additionally generate textual explanations using Qwen2-VL. To generate clinically meaningful explanations, a structured prompt is constructed using the generated caption and the grounded findings obtained from the visual grounding stage. The prompt explicitly constrains the language model to explain only the visual evidence identified by CLIPSeg and discourages unsupported diagnostic conclusions.

The prompt follows the template:

Explainable Medical Captioning Prompt

You are assisting in explainable medical image captioning.
Caption: *<generated caption>*
Grounded visual findings: *<grounded phrases>*
Generate exactly this format:
Visual evidence:
Clinical interpretation:
Counterfactual reasoning:
Limitation:
Rules:
Do not invent new diseases.
Explain only the caption and grounded phrases.
Be concise.

This structured prompting strategy encourages faithful explanations while reducing hallucinated findings. The generated explanation is subsequently integrated into the final explainability panel.

3.3.7. Explainability Panel Generation

The final explainability output is presented as a multi-panel visualisation. The panel contains the original image with phrase-level localisation boxes, the aggregated evidence heatmap, the generated counterfactual image, the corresponding difference map, the clinical explanation, and a colour-coded caption highlighting grounded phrases.

This combined representation enables both visual and textual inspection of the evidence supporting the generated caption and provides a more transparent explanation of the model's prediction.

4. Results

This section presents the qualitative and quantitative results obtained in all the three subtasks across different evaluation metrics mentioned in the overview paper of the task [7].

4.1. Concept Detection

Table 1 presents the official results obtained on the concept detection task. Our retrieval-based framework achieved an F1-score of 0.5234 on the official test set. The system leveraged DenseNet121 embeddings, cosine-similarity retrieval, and adaptive thresholding to predict UMLS concepts associated with medical images.

Although retrieval-based methods provide strong interpretability and computational efficiency, the performance gap relative to the leading approaches suggests that more powerful multimodal representations and concept-aware learning strategies could further improve concept prediction performance.

For the synthetic concept detection track, our method achieved an F1-score of 0.4741, securing the second position among all participating teams.

Table 1
Official concept detection results.

Team	F1 Score	Rank
dsgt_caption	0.5790	1
archimedes	0.5781	2
UIT_Worker	0.5725	3
krishnatewari	**0.5234**	**10**

Table 2
Official synthetic concept detection results.

Team	F1 Score	Rank
AUEB/DMST	0.5609	1
krishnatewari	**0.4741**	**2**
NUSTPB	0.1048	3

The stronger performance on synthetic data suggests that the retrieval framework generalizes reasonably well when image-concept relationships are more structured and less affected by the variability observed in real-world radiology images.

4.2. Caption Prediction

Table 3 summarizes the caption prediction results. Our encoder-decoder captioning framework achieved an overall score of 0.2685, ranking 11th among 13 participating teams. A detailed breakdown of the obtained captioning metrics is presented in Table 4.

Table 3
Official caption prediction results.

Team	Overall	Relevance	Factuality	Rank
AIstatLab	0.3755	0.5786	0.1724	1
UFPE-Essex-UNED	0.3603	0.5527	0.1679	2
dsgt_caption	0.3571	0.5365	0.1777	3
krishnatewari	**0.2685**	**0.4346**	**0.1025**	**11**

Table 4
Detailed caption prediction metrics.

Metric	Score
BERTScore	0.5398
ROUGE-1	0.2059
Similarity	0.6788
BLEURT	0.3137
MedCAT	0.1219
AlignScore	0.0831

The results indicate that the generated captions exhibit moderate semantic similarity with the reference descriptions, as reflected by the BERTScore of 0.5398. However, factuality-oriented metrics such as MedCAT and AlignScore remain comparatively lower, suggesting that while the model captures general visual semantics, it occasionally struggles to generate clinically precise descriptions. This observation highlights the importance of incorporating stronger medical knowledge and tighter integration between concept detection and caption generation modules.

For the synthetic caption prediction track, the proposed framework achieved an overall score of 0.4341, obtaining the fifth position among seven participating teams.

Table 5
Official synthetic caption prediction results.

Team	Overall Score	Rank
AIstatLab	0.5840	1
AUEB/DMST	0.5136	2
UMUTeam	0.4632	3
krishnatewari	**0.4341**	**5**

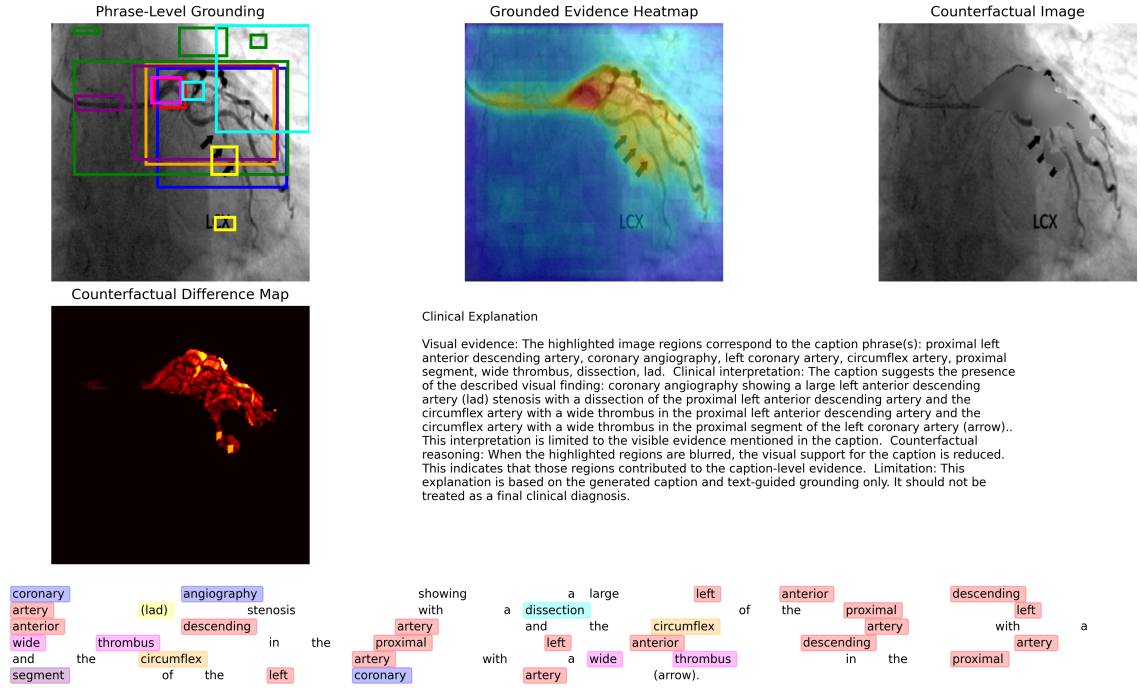

Figure 4: Sample output from the explanation generation pipeline.

4.3. Explainability Results

The official explainability evaluation results are reported in Table 6. Our explainability framework achieved an overall score of 2.85, securing the second position among all participating teams.

Table 6
Official explainability task results.

Team	Read.	Acc.	Detail	Focus-C	Cons.	Compr.	Focus-V	Method	Overall
Morgan State University	4.60	3.60	3.40	4.30	3.10	2.90	3.40	3.00	3.69
krishnatewari	**2.63**	**2.13**	**2.75**	**2.63**	**2.63**	**3.13**	**2.75**	**4.00**	**2.85**
gfreitas	4.50	3.10	3.60	4.10	1.00	2.10	1.00	3.00	2.83

The explainability framework demonstrated competitive performance and ranked second overall. Notably, the proposed method achieved the highest score in methodological appropriateness, indicating that the combination of phrase grounding, visual evidence localization, counterfactual analysis, and explanation generation provides a principled and clinically meaningful explanation pipeline. A sample output produced is shown is Figure 4.

5. Conclusion and Future Work

In this paper, we presented our participation in all three subtasks of the ImageCLEFmedical Caption 2026 challenge: concept detection, caption prediction, and explainability. Our retrieval-based concept

detection framework achieved competitive performance, securing 10th place on the official track and 2nd place on the synthetic track. The encoder-decoder captioning model generated semantically meaningful captions but showed limitations in clinical factuality, highlighting the challenges of medical image caption generation. For the explainability task, our CLIPSeg-based grounding and counterfactual reasoning framework achieved 2nd place overall and received the highest score for methodological appropriateness, demonstrating the effectiveness of combining visual grounding with textual explanations.

Although the lightweight CNN-LSTM architecture provided a reliable and computationally efficient baseline for the challenge, future research may explore large medical vision-language foundation models such as Florence-2, MedCLIP, BioViL-T, and Qwen2.5-VL to improve multimodal representation learning for both concept detection and caption generation. Investigating concept-guided captioning, retrieval-augmented generation, and knowledge-aware decoding could further enhance clinical accuracy and reduce hallucinations. For explainability, future research directions include developing faithfulness-aware evaluation methods, integrating attribution techniques with visual grounding, and exploring causal and counterfactual reasoning frameworks. Additionally, unified multimodal architectures that jointly learn concept detection, caption generation, and explanation generation may provide a promising path toward more accurate, interpretable, and trustworthy medical AI systems.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.4, Claude Sonnet-4.6, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135. doi:https://doi.org/10.1016/j.media.2017.07.005.

[2] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (2019) e1312.

[3] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56.

[4] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia Retrieval in Medical Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[5] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International

Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[6] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, 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 LNCS, Jena, Germany, 2026.

[7] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryılmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – Medical Concept Detection and Caption Generation with Synthetic Data Extensions, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[8] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2097–2106. doi:10.1109/CVPR.2017.369.

[9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 590–597. URL: https://ojs.aaai.org/index.php/AAAI/article/view/3834. doi:10.1609/aaai.v33i01.3301590.

[10] S.-C. Huang, L. Shen, M. P. Lungren, S. Yeung, GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3942–3951. doi:10.1109/ICCV48922.2021.00391.

[11] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, O. Oktay, Making thenbsp;most ofnbsp;text semantics tonbsp;improve biomedical vision–language processing, in: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, Springer-Verlag, Berlin, Heidelberg, 2022, p. 1–21. URL: https://doi.org/10.1007/978-3-031-20059-5_1. doi:10.1007/978-3-031-20059-5_1.

[12] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive learning from unpaired medical images and text, Proc. Conf. Empir. Methods Nat. Lang. Process. 2022 (2022) 3876–3887.

[13] A. Chatzipapadopoulou, I. Pantelidis, F. Charalampakos, M. Samprovalaki, G. Moschovis, P. Kaliosis, K. V. Dalakleidi, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_186.pdf.

[14] D. Mamdouh, M. Attia, M. Osama, N. Mohamed, A. Lotfy, T. Arafa, E. A. Rashed, G. Khoriba, Advancements in radiology report generation: A comprehensive analysis, Bioengineering (Basel) 12 (2025) 693.

[15] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest x-ray report generation, in: F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference,

volume 106 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 249–269. URL: https://proceedings.mlr.press/v106/liu19a.html.

[16] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900. URL: https://proceedings.mlr.press/v162/li22n.html.

[17] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[18] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millicah, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[19] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00461.

[20] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: training a large language-and-vision assistant for biomedicine in one day, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[21] Qwen Team, Qwen2.5: A New Generation of Large Language Models, https://qwenlm.github.io/, 2025. Technical Report.

[22] K. Tewari, S. Pal, Advancing Vision and Language in GI Diagnosis: Florence2 for Question Answering and Stable Diffusion for Image Synthesis, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_208.pdf.

[23] K. Tewari, S. Pal, X-VQA for GI Diagnostics: Multimodal Visual Question Answering with Confidence-Aware Explanations, in: Proceedings of the MediaEval 2025 Multimedia Benchmark Workshop, 2025. URL: https://2025.multimediaeval.com/paper38.pdf.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[25] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.

[26] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[27] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[29] T. Lüddecke, A. Ecker, Image segmentation using text and image prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7086–7096. doi:10.1109/CVPR52688.2022.00695.

[30] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harvard Journal of Law & Technology 31 (2018) 841–887. URL: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf. doi:10.2139/ssrn.3063289.

[31] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2376–2384. URL: https://proceedings.mlr.press/v97/goyal19a.html.

[32] S. Mertes, T. Huber, K. Weitz, A. Heimerl, E. André, Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning, Frontiers in Artificial Intelligence Volume 5 - 2022 (2022). URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.825565. doi:10.3389/frai.2022.825565.

[33] K. Tewari, A. Verma, S. Pal, IReL, IIT(BHU) at MEDIQA-MAGIC 2025: Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_207.pdf.

[34] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, D. Neal, Q. M. Rashid, M. Schaekermann, A. Wang, D. Dash, J. H. Chen, N. H. Shah, S. Lachgar, P. A. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Agüera Y Arcas, N. Tomašev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. K. Barral, D. R. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, V. Natarajan, Toward expert-level medical question answering with large language models, Nat. Med. 31 (2025) 943–950.

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

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.

[38] T. Dettmers, Bitsandbytes: 8-bit Optimizers and Quantization Routines for Deep Learning, https://github.com/bitsandbytes-foundation/bitsandbytes, 2022. GitHub repository.