AUEB NLP Group at ImageCLEFmedical Caption 2025

Notebook for the AUEB NLP Group and Archimedes Unit at ImageCLEFmedical Caption 2025

Anna Chatzipapadopoulou[1,2,*,†], Ippokratis Pantelidis[1,2,*,†], Foivos Charalampakos[1,†], Marina Samprovalaki[1,†], Georgios Moschovis[1,2,*], Panagiotis Kaliosis[3], Kalliopi V. Dalakleidi[1], John Pavlopoulos[1,2] and Ion Androutsopoulos[1,2]

[1]*Department of Informatics, Athens University of Economics and Business, 76, Patission Street, GR-104 34 Athens, Greece*
[2]*Archimedes Unit, Athena Research Center, 1, Artemidos Street, GR-151 25 Athens, Greece*
[3]*Department of Computer Science, Stony Brook University, NY 11794-2424, Stony Brook, USA*

Abstract

This article presents the methodology and results of AUEB NLP Group's and Archimedes Unit's participation in the 9th edition of the ImageCLEFmedical Caption evaluation campaign, addressing the Concept Detection, Caption Prediction, and Explainability tasks. The Concept Detection task involves the automatic association of biomedical images with relevant medical concepts, while the Caption Prediction task focuses on generating clinically meaningful diagnostic captions based on the content of these images. Building upon our previous work, we experimented extensively with image encoders based on Convolutional Neural Networks (CNNs) in combination with Feed-Forward Neural Network (FFNN) classifiers and ensemble approaches. To improve robustness and generalization, we developed diverse ensemble strategies that combine predictions across multiple architectures. Additionally, we applied a per-label thresholding method during inference, allowing the system to fine-tune decision boundaries for each concept individually. For the Caption Prediction task, we used InstructBLIP as the backbone of our pipeline to generate initial captions, which we then refined using a series of enhancement strategies. These included a retrieval-augmented Synthesizer that incorporates information from similar training images, a Multisynthesizer that additionally integrates concept predictions, and LM-Fuser, a lightweight model trained to combine multiple caption hypotheses. Furthermore, we applied the Distance from Median Maximum Concept Similarity (DMMCS) method to guide decoding toward concept-aware captions and used MedCLIP-based re-ranking to further improve visual-textual alignment. We also experimented with reinforcement learning via a mixed training objective that combines cross-entropy and task-specific rewards. In the Explainability task, we generated visual explanations by identifying and localizing key medical entities in the images using a structured prompting approach with GPT-4o. This involved extracting medical terms from generated captions and drawing bounding boxes to connect these terms to visual regions, thereby enhancing clinical decision transparency. Overall, our group ranked 1st in Concept Detection, 5th in Caption Prediction, and 1st in the Explainability task.

Keywords

Natural Language Processing, Computer Vision, Biomedical Images, Convolutional Neural Networks, Multi-Label Classification, Caption Generation, Generative Models, Transformers, Deep Learning, Vision-Language Models, Explainable Artificial Intelligence

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
ann.chatzipapadopoul@aueb.gr (A. Chatzipapadopoulou); ippokratispantelidis@gmail.com (I. Pantelidis); phoebuschar@aueb.gr (F. Charalampakos); mar.samprovalaki@aueb.gr (M. Samprovalaki); geomos@aueb.gr (G. Moschovis); pkaliosis@cs.stonybrook.edu (P. Kaliosis); dalakleidi@aueb.gr (K. V. Dalakleidi); annis@aueb.gr (J. Pavlopoulos); ion@aueb.gr (I. Androutsopoulos)
https://www.linkedin.com/in/anna-chatzipapadopoulou/ (A. Chatzipapadopoulou); https://www.linkedin.com/in/ippokratis-pantelidis/ (I. Pantelidis); http://www.linkedin.com/in/marina-samprovalaki/ (M. Samprovalaki); https://geomos.sites.aueb.gr/ (G. Moschovis); https://pkaliosis.github.io (P. Kaliosis); https://kvdalakleidi.wordpress.com/ (K. V. Dalakleidi); https://ipavlopoulos.github.io/ (J. Pavlopoulos); https://www.aueb.gr/users/ion/ (I. Androutsopoulos)
0009-0005-1633-4966 (A. Chatzipapadopoulou); 0009-0006-9528-5505 (I. Pantelidis); 0000-0003-0547-0581 (G. Moschovis); 0000-0001-6702-4398 (K. V. Dalakleidi); 0000-0001-9188-7425 (J. Pavlopoulos)

1. Introduction

ImageCLEF [1] is an ongoing evaluation initiative, first launched in 2003 under the Conference and Labs of the Evaluation Forum (CLEF)[1], with the goal of promoting the development and benchmarking of technologies for annotation, indexing, classification, and retrieval across multi-modal data. One of the central tracks in the campaign is ImageCLEFmedical, which focuses on real-world medical imaging challenges.

This year marked the 9th edition of the ImageCLEFmedical Caption task [2], where we participated in all three tasks: (i) **Concept Detection**, which aims to automatically associate medical images with relevant biomedical concepts (tags); (ii) **Caption Prediction**, which focuses on generating concise, accurate diagnostic descriptions based on medical image content; and (iii) the newly introduced **Explainability** task, which is about supporting clinicians in building trust in black-box models.

Concept Detection aids the interpretation of medical images by identifying relevant biomedical concepts, while Caption Prediction focuses on generating diagnostic summaries that describe visual findings and anatomical structures. Rather than replacing clinicians, the systems developed for these tasks are designed to support the diagnostic process by highlighting key image regions, accelerating reporting, and reducing the risk of missed information. When used effectively, they can improve both the speed and consistency of medical assessments [3]. The Explainability task reflects the increasing emphasis on explainability in medical Deep Learning (DL) based systems. Linking textual outputs to visual evidence —such as bounding boxes around referenced entities— promotes transparency and can help build user trust in clinical settings, especially for safety-critical decisions.

1.1. AUEB NLP Group and Archimedes Unit Contributions

This paper presents the methods and experimental systems developed by the AUEB NLP Group and Archimedes Unit for the 2025 editions of the Concept Detection, Caption Prediction, and Explainability tasks of the ImageCLEFmedical challenge [2]. Our approaches leverage recent advances in multimodal AI, particularly instruction-tuned Large Language Models (LLMs) [4], which drove both our captioning strategies and our generation of visual explanations in the Explainability task.

Our submission to the Concept Detection task focuses on a method that combines visual feature extraction with concept classification. We used a Convolutional Neural Network (CNN) encoder to extract visual features from the medical images. These features were fed into a Feed-Forward Neural Network (FFNN) to classify the images into various medical concepts. We experimented with a range of CNN backbones, including EfficientNet-B0, DenseNet, and ConvNeXt, to assess the impact of different architectural choices on predictive performance. To improve robustness, we explored various ensembling techniques. These included ensembles of models using union- and intersection-based aggregation strategies. Our final submissions featured both individual models and ensembles.

Regarding the Caption Prediction task, our methodology comprised seven main approaches. The first approach was a fine-tuned InstructBLIP model [5] trained on the extended version of the Radiology Objects in Context Version 2 (ROCOv2) dataset [6], which served as the baseline for generating initial captions. Most of the remaining approaches built upon this baseline, refining its output through a series of downstream strategies aimed at enhancing clinical accuracy, fluency, and alignment with visual content. The second approach, the *Synthesizer*, employed retrieval-augmented generation: visually similar training images were identified based on embedding proximity, and their captions were fed, along with the test image, into a Visual Language Model (VLM) such as Idefics2 to refine the output [7]. The *Multisynthesizer* extended this by incorporating Unified Medical Language System (UMLS) [2] concepts predicted by our Concept Detection system into the prompt, enhancing domain-specific accuracy. The fourth approach employed the *Distance from Median Maximum Concept Similarity (DMMCS)* algorithm [8] to guide decoding toward concept-aware captions, biasing generation toward clinically relevant terms. In the fifth approach, we introduced *LM-Fuser* [9], a lightweight FLAN-T5 model [10]

[1] https://www.clef-initiative.eu/, Last accessed: 2025-05-23

[2] UMLS: https://www.nlm.nih.gov/research/umls/index.html, Last accessed: 2025-05-20

trained to fuse multiple candidate captions into a single coherent output by leveraging their complementary strengths. Our sixth approach incorporated *MedCLIP* [11] as a test-time reranker, selecting among multiple beam-generated captions based on vision-language similarity, thereby improving visual grounding and reducing hallucinations. Finally, we developed the *Mixer* framework, which applied reinforcement learning via Self-Critical Sequence Training (SCST) [12] to optimize a mixed objective combining cross-entropy loss with evaluation-based rewards, such as BERTScore, ROUGE, BLEURT, UMLS F1, and AlignScore.

Building on our track record of successful participation in the ImageCLEFmedical campaign [13, 14, 15, 16, 17, 18], the AUEB NLP Group and Archimedes Unit submitted systems to all three tasks of the ImageCLEFmedical Caption 2025 edition. Our submissions achieved 1st place in the Concept Detection task out of 9 participating teams, 5th place in the Caption Prediction task among 8 teams, and 1st place in the newly introduced Explainability task, which included 2 participating teams. §2 provides an overview of this year's dataset, while §3 describes the methodologies employed for each task. In §4, we report our experimental results and performance metrics. Finally, §5 concludes the paper with a summary of our findings and directions for future work. All code used for our experiments is available on GitHub.[3]

2. Data

In this year's edition of the ImageCLEFmedical Caption task, the dataset is composed of radiology images sourced from biomedical articles of the PubMed Central Open Access (PMC OA) subset.[4] It is based on an extended version of the ROCOv2 dataset [6], incorporating additional images and updated annotations. This extended dataset serves as the foundation for all three tasks in the 2025 challenge: Concept Detection, Caption Prediction, and Explainability.

The full dataset initially comprised 97,368 radiology images, each annotated with one diagnostic caption and a set of medical concepts expressed as UMLS Concept Unique Identifiers (CUIs). The organizers provided a predefined split consisting of 80,091 images for training and 17,277 for validation. To facilitate internal evaluation and parameter tuning, we combined the two official subsets and re-partitioned the data into three new subsets: training, validation, and development.

Our stratified re-splitting was conducted to preserve the statistical distribution of the data in terms of both the CUIs and the length of the captions. We adopted a 75%–10%–15% ratio for training, validation, and development, respectively. This resulted in 73,027 images allocated for training, 9,736 for validation, and 14,605 for the development set. All our system variants were evaluated on these internal subsets during development, while final submissions were assessed on the hidden official test set.

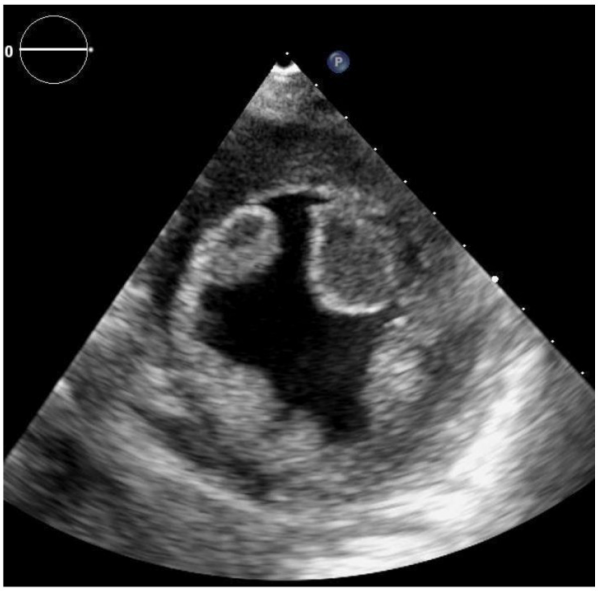
The official test set for 2025 includes 19,267 previously unseen radiology images from ROCOv2 [6], which serves as the benchmark for comparative evaluation across all participating systems.

2.1. Concept Detection

Concept Detection is a multi-label classification problem encompassing 2,479 distinct biomedical concepts derived from UMLS [19]. In this task, the objective is to accurately identify and assign relevant medical concepts (tags) depicted in each image, such as specific medical conditions or procedures. The complete set of concepts includes various modalities of medical imaging, notably X-Ray Computed Tomography, Ultrasonography, Magnetic Resonance Imaging (MRI), and Positron Emission Tomography/Computed Tomography (PET/CT) scans. Each concept is uniquely represented by a CUI following the UMLS standard. A representative example of an image along with its corresponding ground truth concepts is presented in Figure 1.

The distribution of these biomedical concepts exhibits a pronounced imbalance, characterized by a long-tail distribution as seen in Figure 2. Certain concepts are exceptionally frequent, appearing in

[3] https://github.com/nlpaueb/imageclef2025, Last accessed: 2025-05-30.
[4] PMC Open Access: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, Last accessed: 2025-05-20



CUI	UMLS Term
C0041618	Ultrasonography
C0225897	Left ventricular structure
C0030352	Structure of papillary muscle
ID: ImageCLEFmedical_Caption_2025_train_41494	
CC BY [Magdás et al. (2021)]	

Figure 1: This figure, under CC BY from Magdás et al. (2021), presents an example from the ImageCLEFmedical 2025 dataset [6], illustrating the corresponding Concept Unique Identifiers (CUIs) and Unified Medical Language System (UMLS) terms.

more than 34,000 images, whereas many other concepts are exceedingly rare, associated with only a single image each. Table 1 lists the ten most frequently occurring concepts in the ImageCLEFmedical 2025 dataset [6], predominantly corresponding to general medical imaging examinations such as X-Ray Computed Tomography and Plain X-ray. Typically, images contain at least one of these overarching medical imaging modalities, accompanied by additional, more specialized concepts.

Table 1
The ten most frequent concepts (CUIs) of the ImageCLEFmedical 2025 dataset [6], along with their corresponding UMLS terms, and the number of images they are associated with.

Most Common Concepts			
Rank	**CUI**	**UMLS Term**	**Images**
1	C0040405	X-Ray Computed Tomography	34,055
2	C1306645	Plain x-ray	26,531
3	C0024485	Magnetic Resonance Imaging	15,475
4	C0041618	Ultrasonography	14,237
5	C0817096	Chest	12,559
6	C0002978	angiogram	5,387
7	C0000726	Abdomen	5,300
8	C0037303	Bone structure of cranium	4,715
9	C0030797	Pelvis	4,449
10	C0023216	Lower Extremity	3,911

Conversely, a substantial portion of the concept set is rarely represented; notably, twelve illustrative rare concepts are presented in Table 2, each appearing in exactly one image. The presence of these rare labels underscores the considerable challenge posed by data sparsity and highlights the complexities inherent in accurately modeling rare but potentially clinically important phenomena.

Our exploratory analysis also reveals notable variation in the number of concepts assigned to individual images. Specifically, the maximum number of concepts assigned to a single image is 28, a case occurring only once, while the minimum number (a single concept per image) occurs in 10,018 images. On average, each image is annotated with approximately 3.20 concepts.

Table 2
Twelve example concepts (CUIs) from the ImageCLEFmedical 2025 dataset [6], each appearing in only one image and presented along with its corresponding UMLS terms.

CUI	UMLS Term
C0598801	Diffusion weighted imaging
C0202657	CT follow-up
C1956110	Cone-Beam Computed Tomography
C0011906	Differential Diagnosis
C0040395	tomography
C1690005	MRI venography
C0243032	Magnetic Resonance Angiography
C0183062	Root canal post
C0203668	Radioisotope scan of bone
C0412650	Computed tomography of cervical spine
C1962945	Radiographic imaging procedure
C0225273	Structure of adductor canal

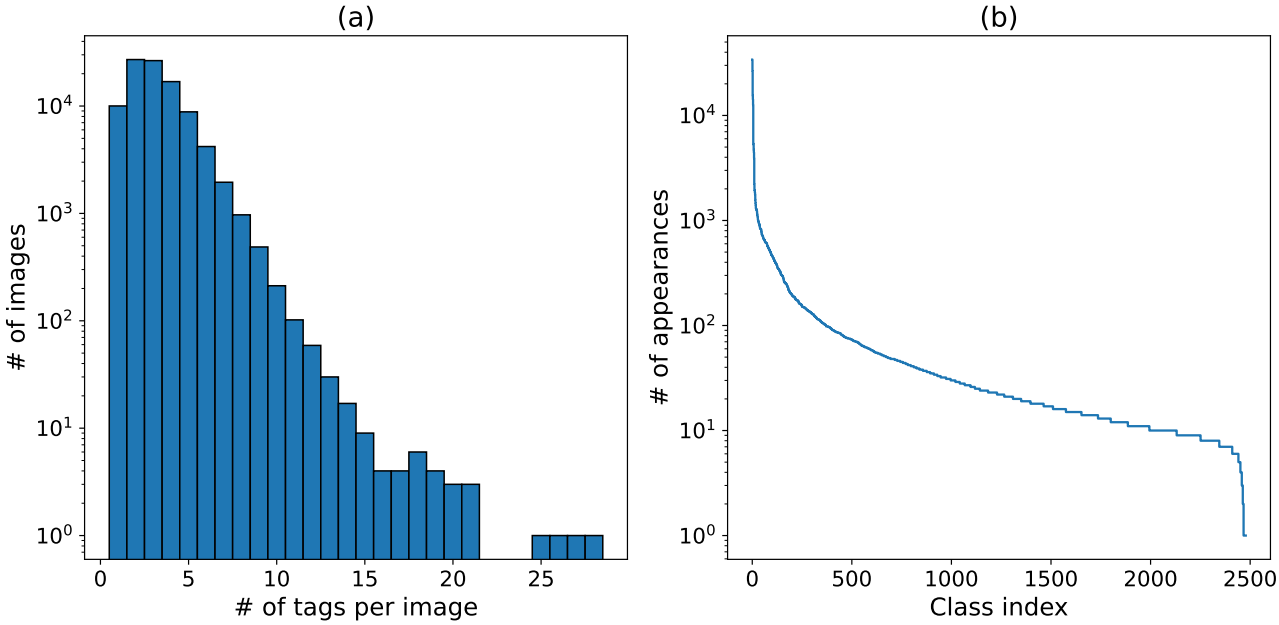


Figure 2: (a) Histogram showing the number of gold concepts per image. (b) Visualization of the dataset's long-tail distribution: the y-axis shows the number of occurrences of each concept, and the x-axis the concept's class index.

2.2. Caption Prediction

Each image in the dataset is paired with a diagnostic caption summarizing the visual medical content. For the 2025 edition, a total of 97,368 captions are provided, one for every image. Among these, 96,866 are unique, corresponding to a uniqueness rate of 99.48%. Captions vary significantly in length: the longest consists of 778 words (appearing once), while the shortest comprises a single word (noted in 81 instances). On average, captions contain 21.04 words.

We ensured that the caption length distribution remains stable across our internal training, validation, and development splits. To facilitate a better understanding of this distribution, Figure 3 visualizes the caption lengths using both histogram and box plot representations. A logarithmic scale is employed on the y-axis to better capture the wide range of caption frequencies and highlight rare outliers.

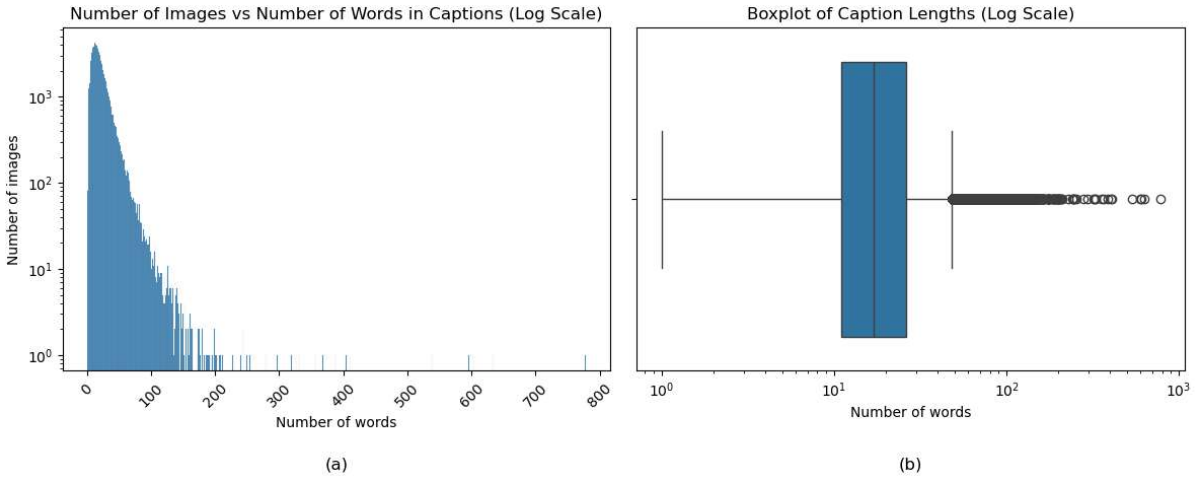


Figure 3: (a) Histogram showing the distribution of caption lengths in the 2025 dataset. The y-axis uses a logarithmic scale to accommodate the skewed distribution. (b) Corresponding box plot highlighting the spread and outliers of the same distribution.

Despite the high uniqueness of the captions, certain phrasing patterns recur, typically reflecting routine imaging procedures. Table 3 lists the five most frequently observed captions, most of which refer to panoramic or chest radiographs. Meanwhile, the most frequent non-trivial words—excluding stopwords—are summarized in Table 4. Common terms include medical imaging modality indicators (e.g., *ct*, *tomography*), laterality markers (*right*, *left*), and descriptive verbs like *showing* and *shows*, underscoring the consistent linguistic patterns across diagnostic reports.

According to the task organizers, all captions are subjected to a pre-processing pipeline prior to evaluation. Specifically:

- All characters are converted to lower-case.
- Numerical values are normalized into their word equivalents (e.g., “10” becomes “ten”).
- Punctuation marks are removed.

Table 3
Most frequently occurring gold captions in the ImageCLEFmedical 2025 dataset [6], along with their respective frequencies.

Most common captions		
Rank	**Caption**	**Occurrences**
1	Initial panoramic radiograph.	41
2	Final panoramic radiograph.	37
3	Chest X-ray.	32
4	Chest radiograph.	17
5	Pretreatment panoramic radiograph.	10

2.3. Explainability Task

The Explainability Task involved 16 radiology images selected from the official test set. Only the raw, unannotated images were provided to participants—no diagnostic captions, concept labels, or metadata

were included. Participants were asked to generate visual explanations for captions that they themselves produced using their own captioning models. These visual explanations consisted of bounding boxes that localize and ground specific medical terms or phrases from the generated captions directly onto the image. There were no restrictions on the explanation format or method, and participants were encouraged to be creative.

Table 4
Top ten most frequent non-stopword tokens in gold captions of the ImageCLEFmedical 2025 dataset [6].

Most frequent content words (excluding stopwords)	
Word	**Count**
showing	27,418
arrow	23,065
right	22,782
left	22,622
ct	21,360
image	14,109
chest	12,865
tomography	12,489
scan	12,266
computed	11,184

3. Methods

This section outlines the methods employed in our submissions to the Concept Detection, Caption Prediction, and Explainability tasks.

3.1. Concept Detection

Building upon our prior research [13, 14, 15, 16, 20], our submissions for this year's Concept Detection task were based on classification models composed of neural image encoders. Furthermore, we submitted several ensemble systems that employed strategies such as union-based and intersection-based aggregation.

3.1.1. CNN-FFNN

Our main system employs a Convolutional Neural Network (CNN) as the primary backbone for feature extraction, coupled with a Feed-Forward Neural Network (FFNN) serving as the classification module. Specifically, the CNN backbone generates spatially structured feature maps from the input images. To derive a single image embedding per image, we apply global Generalized-Mean (GeM) pooling [21] which incorporates learnable pooling parameters, allowing it to include traditional pooling mechanisms such as (global) max pooling and (global) average pooling as special cases.

The FFNN classifier consists of an output layer with $|C|$ neurons, each neuron corresponding to a concept in the dataset. Each neuron uses a sigmoid activation function, converting the raw logits into probabilities. A concept is assigned to an image if its predicted probability surpasses a (global) threshold τ. The threshold value was determined through grid search, optimizing the primary evaluation metric (F1-score) on the validation set.

The system is trained by minimizing the binary cross-entropy loss, treating each concept as an independent binary classification task and summing the resulting losses. Optimization is performed using the Adam optimizer [22] with a learning rate of 1e-3. A learning rate decay schedule is applied, reducing the learning rate upon plateauing of the validation loss with a patience of one epoch. Early stopping is employed based on validation loss, with a patience of three epochs to prevent overfitting.

The models are trained for up to 100 epochs with a batch size of 16. Input images are resized and normalized. All models are initialized from ImageNet-pretrained weights to leverage transfer learning.

In order to form the ensembles, we trained several instances of this system, experimenting with several image encoders and using different random initializations, and combined them using the UNION and the INTERSECTION of their predicted concept sets. More details about our submitted ensemble systems can be found in subsection 3.1.3.

3.1.2. Per-label Threshold Optimization

Given the multi-label nature of the task, apart from the single, global threshold τ (§3.1.1), we also experimented with a second strategy that learned an *individual* threshold τ_c for every concept $c \in \{1, \ldots, C\}$. Let $\boldsymbol{S} \in [0,1]^{N \times C}$ denote the validation set's prediction score matrix returned by a CNN + FFNN model, with $\boldsymbol{S}_{ic}$ being the probability of concept c for sample i (N is the number of samples). The objective is to maximize the main evaluation metric of the Concept Detection task which is the samples-average F_1. More specifically, the score was computed by averaging the individual F_1 scores over all images (in the corresponding set). For each image t in the set T, an individual F_1 score $\hat{f}_1$ was calculated based on the overlap between the predicted concept set p_t and the ground truth set g_t, both represented as binary multi-hot vectors. The final (global) F_1 score, denoted as F_1, was then obtained by averaging the individual scores across all images in the set.

$$F_1 = \frac{1}{|T|} \sum_{t \in T} \hat{f}_1(p_t, g_t) \tag{1}$$

Because this metric considers all the labels of a given image, the thresholds cannot be optimized independently. We therefore employ the **coordinate-ascent algorithm**, detailed below:

Initialization: start from an initial vector $\boldsymbol{\tau}^{(0)}$
One pass over all concepts:
For each concept c:

a) sort the N scores of column c in descending order, obtaining $\boldsymbol{S}_{1c} \geq \ldots \geq \boldsymbol{S}_{Nc}$
b) for $k = 1, \ldots, N$ *tentatively* set $\tau_c = \boldsymbol{S}_{kc}$, flip samples with $\boldsymbol{S}_{ic} \geq \tau_c$, recompute the global F_1, and keep the best-achieved value $F_1^\star$ together with its threshold $\tau_c^\star$
c) if $F_1^\star$ exceeds the current global F_1, accept the update $\tau_c \leftarrow \tau_c^\star$; otherwise leave τ_c unchanged

Stopping criterion: repeat the pass until a full sweep over $c = 1, \ldots, C$ makes no further improvement (empirically, two to three passes in our experiments).

3.1.3. Ensemble Strategies

To enhance robustness and predictive accuracy, we developed a range of ensemble strategies that combined predictions from models trained with diverse configurations and architectures. These ensembles were formulated both at the model level—through variation in architectures—and at the prediction level—by aggregating multiple prediction outputs.

Our ensemble experiments involved models trained using three distinct CNN encoders: **EfficientNet-B0** [23], **DenseNet-121** [24], and **ConvNeXt-Tiny** [25]. Specifically for the EfficientNet-B0 encoder, we conducted an ensembling approach based on Monte-Carlo cross-validation. This method involved creating five different train-validation splits from the original dataset. For each of these splits, we trained a different classifier equipped with an EfficientNet-B0 encoder, maintaining a consistent development set across all splits. During inference, each of the five trained models produced individual prediction outputs, which we aggregated using the intersection operation, retaining only those concepts predicted by all models. This aggregated prediction set was subsequently combined (via union) with predictions from an additional EfficientNet-B0 model trained on the entire available training and validation set, as well as with the predictions from DenseNet-121 and ConvNeXt-Tiny.

In addition to basic union and intersection operations, we explored two advanced aggregation strategies to provide more nuanced concept inclusion:

- **Dual Threshold Aggregation**: To balance high precision with improved recall, we implemented a dual-threshold aggregation strategy based on model-level agreement. Let $V_{i,j}$ denote the number of models that predicted concept j for image i, and let M be the total number of models.
  We first define a *core* set of concepts with full agreement across all models:

$$\text{core}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} = M \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

  To incorporate additional concepts with partial yet substantial consensus, we introduce a *border* set containing concepts predicted by at least L models (with $L < M$):

$$\text{border}_{i,j} = \begin{cases} 1, & \text{if } L \leq V_{i,j} < M \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

  The final prediction for each concept is then determined by the union of the core and border sets:

$$\hat{P}_{i,j} = \text{core}_{i,j} \cup \text{border}_{i,j} \tag{4}$$

  This approach guarantees that highly confident predictions (i.e., full agreement) are always preserved, while still allowing for broader concept coverage when a sufficient level of consensus is observed among models.
- **Partial Intersection Aggregation**: This strategy adopts a hierarchical, consensus-driven approach to concepts' aggregation. For each image i and concept j, as above, $V_{i,j}$ denotes the number of models that assigned concept j to image i and M is the total number of models.
  We compute first the strict intersection across all models (as in Dual Threshold Aggregation):

$$\text{core}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} = M \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

  If the set of predicted concepts for a given image i is non-empty (i.e., $\sum_j \text{core}_{i,j} > 0$), we define the final prediction $\hat{P}_{i,j}$ using only the core:

$$\hat{P}_{i,j} = \text{core}_{i,j} \tag{6}$$

  Otherwise, for images where the intersection is empty (i.e., $\sum_j \text{core}_{i,j} = 0$), we fall back to a relaxed criterion and include concepts predicted by at least L (i.e., 2 or 3) models:

$$\hat{P}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} \geq L \\ 0, & \text{otherwise} \end{cases} \quad \text{for } \sum_j \text{core}_{i,j} = 0 \tag{7}$$

  This fallback mechanism ensures that even in cases of model disagreement, each image still receives a set of concept predictions with partial consensus, while prioritizing precision when full agreement is available.

3.2. Caption Prediction

Our submissions for the Caption Prediction task were primarily built around a finetuned InstructBLIP model [5] (§3.2.1), which served as the foundation for many, though not all, of our systems. We developed several extensions, including synthesizing and multi-synthesizing approaches (§3.2.2 and §3.2.3), an LM-Fuser (§3.2.4), and a guided-decoding method, DMMCS [8] (§3.2.5), which leverages concept tags predicted by our CNN-FFNN (§3.1.1). Additional strategies included a test-time reranker using MedCLIP [11] (§3.2.6) and a reinforcement learning-based training scheme, Mixer (§3.2.7), grounded in Self-Critical Sequence Training [12].

3.2.1. InstructBLIP

InstructBLIP [5] is a general-purpose multimodal model designed for instruction-following tasks involving both visual and textual modalities. It employs instruction tuning [26], a technique that refines model behavior based on explicit natural language prompts, thereby enhancing its controllability and adaptability across diverse tasks. The architecture consists of three key components: a frozen image encoder, a Q-Former [27], and a large language model (LLM). The image encoder generates embeddings from the visual input, which are then processed by the Q-Former to extract instruction-aware features conditioned on the input prompt. These features are subsequently passed to the LLM, which generates coherent and contextually grounded textual descriptions. While InstructBLIP is not inherently specialized for the medical domain, we fine-tuned it on our training set for the caption prediction task. It served as the backbone of our pipeline, providing the base outputs for several of our extended captioning systems.

3.2.2. Synthesizer

The Synthesizer [7] is a retrieval-augmented captioning system designed to improve the quality of image descriptions by leveraging visually similar examples. It is built on the idea that images sharing similar visual features tend to have corresponding captions with similar content [28, 29]. For a given test image, we first compute image embeddings using a CNN-FFNN architecture [20], which was originally developed for Concept Detection (Section 3.1.1). Based on cosine similarity, we retrieve the k most similar images from the entire dataset (training, validation, and development). These neighbors, each paired with their corresponding ground-truth captions, form a pool of auxiliary visual and textual context. We experimented with $k \in \{1, 3, 5\}$ and found that $k = 5$ yielded the best performance on our validation set, consistent with our findings in ImageCLEF2024 [18, 30], and thus used it for all subsequent experiments.

Next, we generate an initial draft caption for the test image using our fine-tuned InstructBLIP model (Section 3.2.1). This draft, together with the retrieved captions and the test image, is then passed to Idefics2 [31, 32], a large multi-modal architecture that includes a vision encoder and cross-attention layers for jointly processing image and text inputs. While we did not modify the Idefics2 model, it is inherently capable of integrating multi-modal cues. This design enables the model to refine the initial caption by combining information from the neighbouring and test image, retrieved captions, and the draft caption [7, 9].

Figure 4 depicts the general process, beginning with the draft caption generated by InstructBLIP, combining it with captions from neighboring images, and then using a VLM—specifically Idefics2—to produce a refined caption [7, 18]. Figure 5 presents a test image alongside an initial caption generated by InstructBLIP [5], a neighboring image with its caption, and the final caption produced by Idefics2. For comparison, the gold caption is provided, with similarities indicated in bold. The initial caption correctly identifies the modality and visual markers (e.g., arrows), while the refined caption integrates information from both the test image and the neighbor's caption, adding details such as congestion and possible pleural effusion—consistent with the gold caption [7, 18].

3.2.3. Multisynthesizer

The Multisynthesizer extends the Synthesizer by including predicted medical concept tags during caption refinement. In addition to the input image, draft caption, and captions from visually similar neighbors, the prompt to the multimodal LLM also includes UMLS-based tags predicted by one of our best image tagger (Section 3.1.1). These tags provide additional semantic context, helping the model produce more accurate and clinically relevant captions.

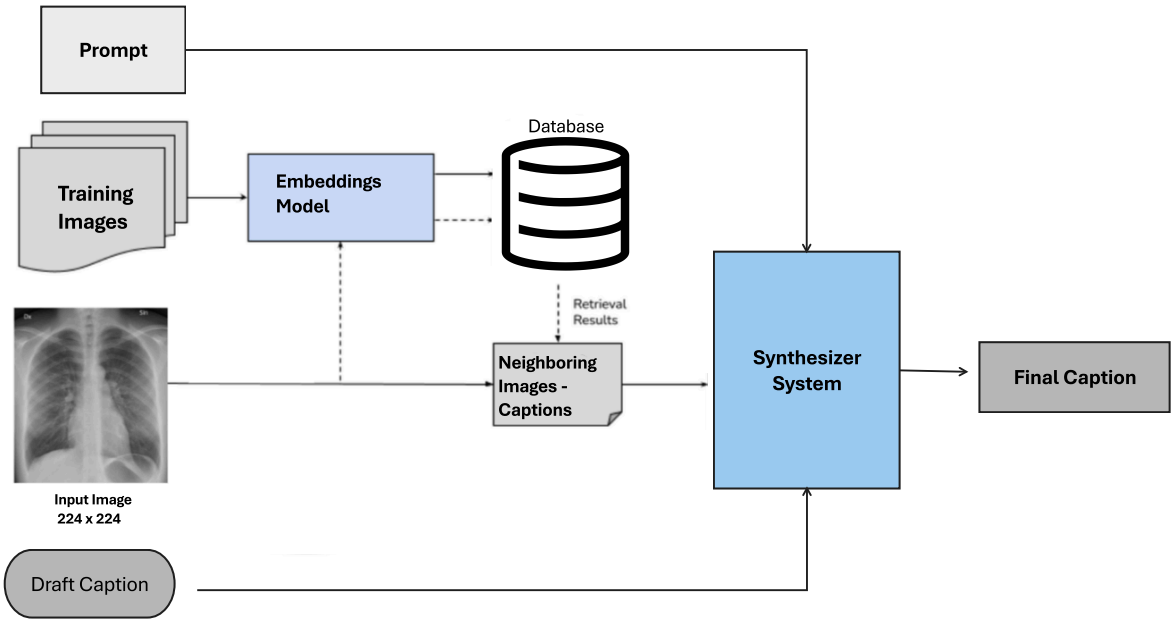


Figure 4: The Synthesizer architecture retrieves neighboring images through their embeddings and uses them together with the embeddings of the input image and a draft caption generated by our fine-tuned InstructBLIP model to produce the final caption guided by the prompt. Figure taken from [7]; the radiology image shown is from the ImageCLEFmedical Caption 2024 dataset [6].

3.2.4. LM-Fuser

LM-Fuser [7, 9] is a caption fusion method designed to improve caption quality by combining predictions from multiple pretrained vision-language models (VLMs) without fine-tuning them individually. Unlike traditional approaches that rely on heavy fine-tuning or resource-intensive few-shot learning, LM-Fuser introduces a more efficient alternative by delegating the fusion task to a smaller language model. Specifically, it uses a Flan-T5 model fine-tuned solely on outputs produced by other captioning systems.

As illustrated in Figure 6, the process begins with a medical image and a task description instructing the generation of a precise and informative caption. Multiple VLMs (namely LLaVA-1.5, LLaMA-3.1, and Idefics2) generate diverse captions for each image in the training and validation sets. These captions, along with the corresponding gold-standard annotation, form the input to the LM-Fuser training dataset.

The LM-Fuser model is trained to map a set of candidate captions to a single, high-quality output caption. During training, the input to Flan-T5 consists of three alternative candidate captions, while the target is the ground-truth caption. Notably, this architecture does not have access to the image itself—it relies purely on text-based input, leveraging the diversity and complementary nature of the predictions. Training was conducted using cross-entropy loss, and ROUGE-L was used as the early stopping criterion due to its computational efficiency.

During inference, the candidate captions are passed to LM-Fuser, which processes their logits—the raw outputs from its decoder—before applying a softmax layer to produce a probability distribution over the vocabulary. Caption generation is then performed using beam search, enabling the model to explore multiple likely sequences before selecting the most coherent one.

By consolidating multiple perspectives from different models, LM-Fuser improves caption reliability without incurring the high computational costs of multimodal fine-tuning. This makes it a practical solution for settings where access to VLM internals is restricted or inference efficiency is paramount.

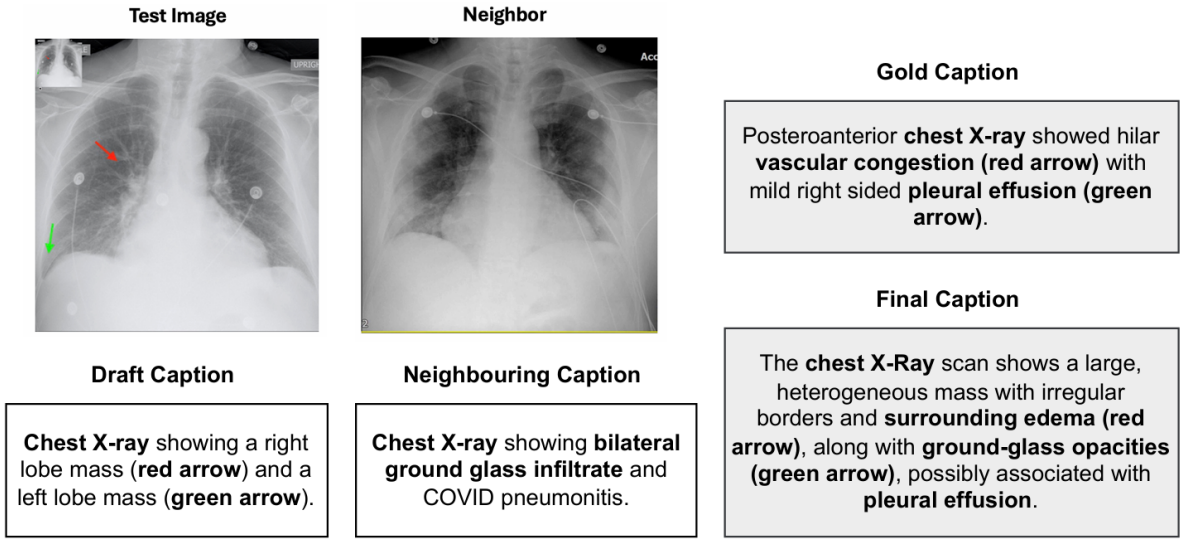


Figure 5: The test image is displayed together with its draft caption, alongside the neighboring image and its corresponding caption. The final caption demonstrates the model's ability to leverage both visual and textual information from the neighboring image. For reference, the gold caption is also provided. All images and the original caption are sourced from [6]. Figure adapted from [7].

3.2.5. DMMCS

We used the DMMCS strategy [8] as a guided decoding method to improve the alignment of generated captions with clinically relevant content. The key idea is to adjust the decoding process based on the predicted medical tags of each image, encouraging the model to include appropriate clinical concepts in its output. DMMCS modifies the scoring function during generation without altering the model architecture or requiring additional training. We applied it on top of several of our models, with the guidance strength controlled by a weighting parameter α. For a detailed explanation of the method, we refer the reader to [8].

3.2.6. MedCLIP Reranker

To enhance caption selection during inference, we implemented a test-time reranking strategy using MedCLIP [11], a contrastive vision-language model pre-trained specifically on radiology data. This method is model-agnostic and can be seamlessly integrated with any captioning backbone, such as InstructBLIP (§3.2.1) or LM-Fuser (§3.2.4).

During inference, instead of generating a single caption per image, we produce a set of $m = 4$ candidate captions via beam search. Each candidate is then scored by MedCLIP, which encodes both the image and the captions into a shared multimodal embedding space. A similarity score is computed between the image and each candidate caption, and the caption with the highest similarity is selected as the final output.

MedCLIP builds upon the CLIP architecture [33], adapting it to the medical domain through contrastive training on paired radiology images and textual reports. The core idea is to learn aligned visual and textual representations that preserve domain-specific semantics. Leveraging this alignment, our reranker prioritizes captions that are not only fluent but also better grounded in the visual evidence. This strategy helps mitigate hallucinations and reinforces the clinical validity of generated descriptions.

3.2.7. Mixer

To better align training objectives with evaluation-time criteria, we implemented a mixed training strategy termed Mixer, which combines cross-entropy loss with reinforcement learning through Self-

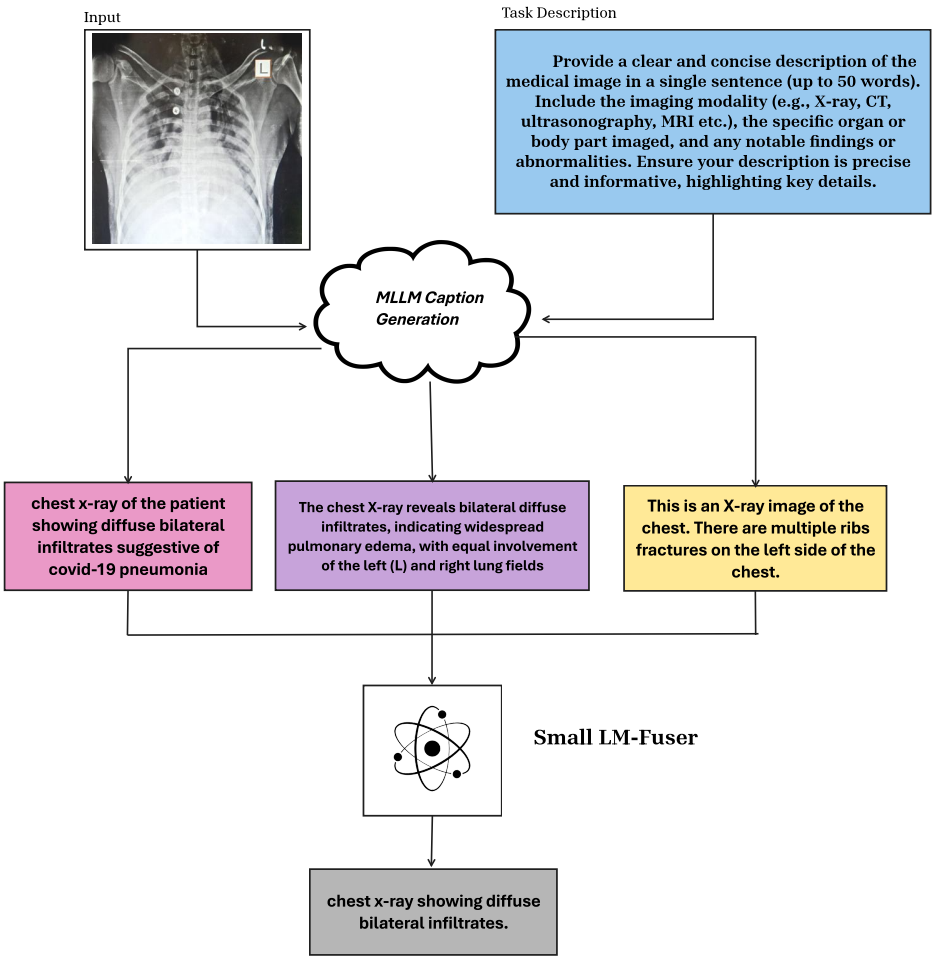


Figure 6: An illustration of LM-Fuser, which combines outputs from multiple MLLMs to produce a more concise caption by integrating diverse candidate predictions. The radiology image used within the figure is sourced from the ImageCLEFmedical Caption 2024 dataset [6]. Figure taken from [7].

Critical Sequence Training (SCST) [12]. This hybrid objective is designed to address exposure bias and directly optimize for evaluation metrics used in the Caption Prediction task.

For each training instance, we generate two types of captions: a *greedy caption* $\hat{y}$, produced via deterministic greedy decoding, and a *sampled caption* y^s, obtained through stochastic decoding (e.g., top-p sampling) or diverse beam search using multiple beam groups. These two candidate captions are evaluated against the gold caption using an internal scoring function that averages multiple task-specific metrics. Specifically, the reward function includes metrics reflecting both relevance (e.g., BERTScore [34], ROUGE-1, BLEURT, and image-text similarity) and factuality (e.g., UMLS Concept F1 and AlignScore), as outlined in the official task definition.

Let $r(y)$ denote the average evaluation score assigned to caption y. The *advantage* of the sampled caption relative to the greedy one is then computed as:

$$Adv(y^s) = r(y^s) - r(\hat{y}), \tag{8}$$

quantifying the improvement (or degradation) of the sampled caption with respect to the greedy baseline under the combined evaluation metric.

Training is guided by a composite loss function that combines the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ with a reinforcement loss $\mathcal{L}_{\text{RL}}$, governed by a mixing coefficient α:

$$\mathcal{L}_{\text{total}} = (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{RL}}. \tag{9}$$

The reinforcement component is computed using the SCST formulation as follows:

$$\mathcal{L}_{\text{RL}} = -Adv(y^s) \cdot \log \pi_\theta(y^s), \tag{10}$$

where $\pi_\theta(y^s)$ is the probability assigned to the sampled caption by the model. This formulation rewards captions that outperform the greedy baseline, while penalizing those that underperform.

To ensure training stability, the reinforcement signal is introduced gradually. Specifically, the mixing coefficient α increases linearly over training epochs. For epoch e out of a total of E epochs, we define:

$$\alpha(e) = \alpha_{\max} \cdot \frac{e+1}{E}, \tag{11}$$

where $\alpha_{\max}$ is the maximum reinforcement weight. This progressive scheduling ensures that early training is dominated by cross-entropy loss—favoring linguistic fluency and stable convergence—while later epochs increasingly emphasize metric-driven optimization aligned with task objectives.

The Mixer approach thus enables end-to-end optimization of captioning models for evaluation-aware performance, without sacrificing the benefits of conventional supervised learning during initial training stages.

3.3. Explainability Task

The newly introduced Explainability Task focuses on enhancing the interpretability of vision-language models by linking textual medical descriptions to specific visual regions within biomedical images. In this task, participants are given only raw radiology images, without any accompanying labels, captions, or metadata. The goal is to produce visual justifications in the form of bounding boxes that correspond to clinically meaningful terms in a caption. These explanations aim to support clinicians in understanding and trusting AI-generated outputs, especially in settings where model decisions are otherwise opaque. Our approach generates these explanations externally, based on predicted captions and medical entity extraction, rather than relying on internal attention mechanisms. While effective in grounding key terms, this strategy does not capture the model's true decision process, limiting its use for full interpretability or causal attribution.

As no ground-truth captions were available and no training was intended for this task, we used InstructBLIP (Section 3.2.1) to automatically generate captions, as it demonstrated the best performance on our held-out development dataset. From these, we extracted medical terms using the domain-specific biomedical NER model `en_core_sci_sm` from the ScispaCy library[5] [35], based on UMLS entities. These entities were subsequently used as targets for bounding box localization. To generate bounding boxes, we explored several prompt engineering strategies using GPT-4o. The most successful prompt adopted a structured, multi-part format designed to balance radiological rigor and linguistic clarity:

- **Preamble:** Introduced the model as a virtual radiology assistant and set expectations for clinically relevant outputs.
- **Input and Task Definition:** Presented the generated caption alongside a concise instruction to draw bounding boxes around image regions corresponding to the detected medical entities.
- **Clarification and Special Cases:** Provided additional rules on handling vague or multi-word concepts, overlapping anatomical regions, and diffuse abnormalities.
- **General Guidelines:** Concluded with emphasis on minimizing hallucinations, ensuring anatomical plausibility, and restricting annotations to observable evidence.

This structured prompting approach led to significantly improved alignment between textual and visual modalities. Figure 7 illustrates a representative output of our system, highlighting the correspondence between identified medical terms and their spatial grounding. The complete prompt used in this task can be found in Appendix 5.

[5] https://allenai.github.io/scispacy/

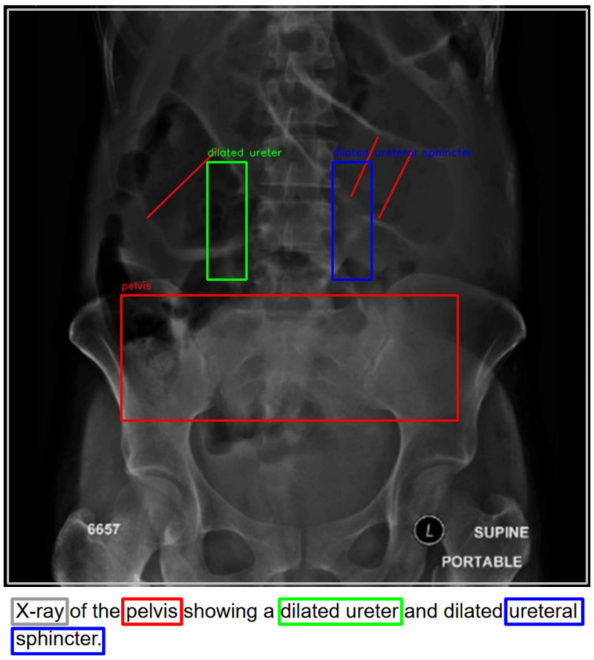


Figure 7: Output of our explainability pipeline on an official test image from ImageCLEFmedical 2025 (`ImageCLEFmedical_Caption_2025_test_118`, CC BY, Muacevic et al. 2024). Key medical terms from the generated caption are linked to predicted bounding boxes.

4. Experiments, Submissions and Results

In this section, we provide details about our experiments regarding this year's evaluation campaign [1]. Moreover, we share details about our submissions and the scores achieved in our held-out development set, as well as the official test set of the competition for both tasks.

4.1. Concept Detection

In the Concept Detection task, we submitted our top 16 models, selected based on performance on our development set, as described in Section (§2). Our submissions included multiple instances of our CNN-FFNN system (§3.1.1), each using different CNN backbones. Specifically, we trained the networks using state-of-the-art CNN architectures, including EfficientNet [23], DenseNet [24] and ConvNextTiny [25]. Moreover, in some of our submissions we incorporated ensemble variants (§3.1.3) that aggregated predictions from these models using union- and intersection-based strategies to improve performance. Additionally, we submitted a model that employed per-label threshold optimization (§3.1.2), in which the decision threshold for each concept was individually tuned via a coordinate-ascent procedure.

As mentioned in Section 3.1.2, our system is evaluated with the F_1 score defined in Eq. (1). Moreover, a secondary evaluation metric (again an F_1 score) was calculated, which only considered manually selected concepts, such as *modality* and *anatomy*.

Our ensemble methods achieved the highest overall performance across both the development and test set [6], outperforming all individual models.

4.2. Additional Concept Detection Experiments

In addition to the models officially submitted in the Concept Detection task, we conducted additional experiments, specifically aimed at enhancing the classification performance for ultrasonography images. Our analysis indicated that our models consistently exhibited lower accuracy for ultrasonography images compared to other modalities, such as X-ray and MRI. To address this issue, we designed

Table 5
Summary of our submissions to the ImageCLEFmedical 2025 Concept Detection task. The table presents the scores of our systems on both our held-out development set and the official test set [6]. It also includes the rankings of these systems among all submissions from the 9 participating teams. **MC**: Monte-Carlo, **EB0**: EfficientNet-B0, **D121**: DenseNet-121, **CN**: ConvNext-Tiny, **INTER**: INTERSECTION, **Dual-L**: number of L models used for thresholding in the ensemble.

Individual Concept Detection Experiments					
Run ID	**Method**	**F1**		**Secondary F1**	**Rank**
		Dev	**Test**		
1980	Dual-3(MC(EB0), D121,CN,EB0)	**0.5973**	**0.5887**	0.9484	1
1981	Dual-3(MC(EB0),D121,EB0)	–	0.5880	0.9506	2
1979	Dual-2(EB0,D121)	–	0.5873	0.9522	3
1977	Dual-2(MC(EB0),EB0)	–	0.5867	0.9449	4
1982	Dual-3(MC(EB0),EB0)	–	0.5866	0.9507	5
1978	Dual-2(MC(EB0))	0.5945	0.5866	0.9465	6
1976	Dual-2(MC(EB0),D121,EB0)	–	0.5864	0.9435	7
1975	Dual-2(MC(EB0),D121,CN,B0)	0.5947	0.5858	0.9388	8
1983	Dual-3(MC(B0))	0.5942	0.5855	0.9515	9
1986	PARTIAL-INTER(MC(EB0),CN,D121)	0.5931	0.5853	**0.9589**	10
1971	CNN-FFNN (EB0)	0.5915	0.5840	0.9488	11
1970	UNION(INTER(MC(EB0)), INTER(EB0,D121,CN))	0.5923	0.5819	0.9520	12
1973	CNN-FFNN (D121)	0.5909	0.5817	0.9462	13
1974	CNN-FFNN (CN)	0.5925	0.5808	0.9334	14
1985	Threshold-per-Label	0.5875	0.5773	0.9456	16
1984	Dual-3(MC(EB0),D121)	0.5954	0.5755	0.9446	20

targeted fine-tuning procedures intended to leverage domain-specific information and improve model accuracy on ultrasonography images.

Initially, we adopted a two-phase fine-tuning strategy. In the first phase, the model was trained on a subset of our training split (henceforth called Dataset 1A) that excluded all ultrasonography images. Upon completion, we preserved both the trained model weights and the associated mapping between output neurons and their corresponding concept labels. This allowed us to retain and correctly position the learned weights when expanding the output layer in the second phase to accommodate the full set of 2,479 labels. In the second phase, we fine-tuned the model on the remaining subset (Dataset 1B) of our training split, consisting exclusively of ultrasonography images. To prepare the model for this phase, the output layer was expanded by adding neurons to accommodate the full set of 2,479 concept labels, since Dataset 1B included additional concepts not present in Dataset 1A. Each neuron corresponds to a specific concept, enabling the model to perform multi-label classification over the complete label set. For labels present in both datasets, the learned weights from the first phase were retained, allowing the model to leverage previously acquired knowledge while adapting more specifically to the new (ultrasonography) modality.

Improving upon this approach, we explored a more advanced masking strategy designed to refine modality-specific predictions. We again partitioned the data from our train split into two subsets: the former excluding ultrasonography images (Dataset 2A), and the latter containing only ultrasonography images (Dataset 2B). A unified set of labels was constructed as the union of labels across both datasets. The model's output layer was structured to accommodate this unified set (i.e., all the available labels of the dataset). During training with Dataset 2A, labels absent from this subset were masked, preventing the model from considering irrelevant label predictions. Similarly, when training on Dataset 2B, labels irrelevant to ultrasonography were masked. By alternating training between these subsets and employing modality-specific label masking, we facilitated effective knowledge transfer while maintaining modality specialization.

These targeted training strategies did not surpass the overall performance achieved by our other models (see tables 5, 11 for comparison). Consequently, these modality-specific models were not selected for final submission. Comprehensive performance details of these supplementary experiments can be found in Appendix 5.

4.3. Caption Prediction

We submitted a total of 26 systems to the Caption Prediction task, leveraging the methods introduced in §3.2. Our submissions span a variety of model combinations and configurations, including base models, guided decoding techniques, reranking strategies, and reinforcement learning.

Several of our systems are based on InstructBLIP (§3.2.1), which served as a foundation for methods such as the Synthesizer and DMMCS (§3.2.2, §3.2.5). We also explored combinations of these components, for example: InstructBLIP paired with DMMCS, or Synthesizer with the MedCLIP Reranker (§3.2.6). The Synthesizer and Multisynthesizer systems were evaluated with two large multimodal backbones—Llama-3.1-8B and Idefics2—with the latter yielding better performance on our development set. The LM-Fuser model (§3.2.4) aggregates predictions from multiple vision-language models—Llama-3.1-8B, Idefics2, and LlaVa-1.5—using a lightweight Flan-T5 model as the fusion layer.

Both the DMMCS guided decoding mechanism and the MedCLIP Reranker were applied on top of all our trained methods, including InstructBLIP and LM-Fuser. Additionally, our Mixer system (§3.2.7) was trained using InstructBLIP as the base model, incorporating reinforcement learning for three epochs. Due to time constraints and computational limitations, training was intentionally kept short.

To provide transparency about the configuration settings of the systems that underwent training, we summarize the key hyperparameters used for InstructBLIP, LM-Fuser, and Mixer in Table 6. These include details on optimizer, learning rate, loss function and batch size.

Table 6
Training configurations for the three trained captioning models.

Model	Loss	Optimizer	LR	Batch Size	Epochs
InstructBLIP	Cross-Entropy	Adam	5e-6	4	38
LM-Fuser	Cross-Entropy	AdamW	1e-4	1	3
Mixer	CE + SCST	Adam	5e-6	1	3

No learning rate scheduler, weight decay, or data augmentations were used for any of the models. The InstructBLIP model was initially configured for 40 training epochs, but early stopping with a patience of 3 halted training at epoch 38 based on the validation loss. The LM-Fuser model was configured to train for 10 epochs. Training was terminated at epoch 3, as the best validation performance, measured by the ROUGE score, was already achieved by that point and showed no further improvement in subsequent evaluations. The Mixer model, due to time constraints, was trained for only 3 epochs. It also employed gradient accumulation, a strategy that allows the model to accumulate gradients over several mini-batches before performing an optimizer step.

To qualitatively illustrate system variation, Table 7 displays captions generated by selected models for the same test image shown in Figure 8. Table 8 summarizes the performance of 11 representative submissions to the ImageCLEFmedical 2025 Caption Prediction task. For each run, we report the Overall, Relevance Average, and Factuality Average scores on both the development and test sets, as well as the corresponding official rank on the test leaderboard.

Table 9 presents a full evaluation of the top-performing submissions across all official test-set metrics, including relevance and factuality sub-averages.

4.4. Explainability Task

To develop our submission for the Explainability task, we used GPT-4o via the OpenAI API [6] to prevent any risk of competition data leakage. Each image was paired with a generated caption (from our captioning pipeline) and a list of extracted medical entities. These were passed to GPT-4o along with a structured system prompt instructing the model to predict bounding boxes corresponding to each term.

[6] https://openai.com/api/

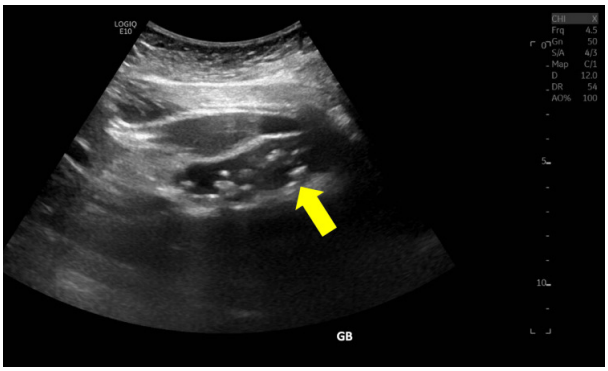


Figure 8: Ultrasound image from the ImageCLEFmedical 2025 dataset [6], used as input for the captions listed in Table 7.

Table 7
Captions generated by selected submitted systems for the test image shown in Figure 8 [6] © [ImageCLEFmedical_Caption_2025_test_18; CC BY, Muacevic et al., 2024].

	Generated Captions
InstructBLIP	Ultrasound image of the right kidney showing a hypoechoic lesion (yellow arrow).
InstructBLIP + DMMCS ($\alpha = 0.1$)	Ultrasound image of the right kidney showing a hypoechoic lesion (yellow arrow).
InstructBLIP + DMMCS ($\alpha = 0.1$) + MedCLIP Reranker	Abdominal ultrasound showing a hypoechoic lesion (yellow arrow) in the left kidney.
LM-Fuser	Abdominal ultrasonography showing a dilated common bile duct (white arrow).
Synthesizer (Idefics2-8B)	Longitudinal view of the right kidney showing an anechoic area in the renal pelvis (arrow) suggestive of hydronephrosis.
Mixer	Transesophageal echocardiography (TEE) of the left ventricle.

Table 8
Performance summary of selected submissions to the ImageCLEFmedical 2025 Caption Prediction task. For each approach, we report the Overall score, the Relevance Average, and the Factuality Average on both our held-out development set (Dev) and the official hidden test set (Test).

Run ID	**Approach**	**Overall**		**Relevance Avg.**		**Factuality Avg.**		**Rank**
		Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**	
1403	InstructBLIP	0.2977	**0.3068**	0.4775	0.4759	0.1180	0.1377	**48**
1463	InstructBLIP + DMMCS ($\alpha = 0.1$)	0.2967	0.3047	**0.4783**	**0.4769**	0.1152	0.1324	50
1724	InstructBLIP + MedCLIP Reranker	**0.2986**	0.3026	0.4740	0.4714	0.1233	0.1339	61
1718	Synthesizer (Idefics2-8B)	0.2876	0.2957	0.4764	0.4733	0.0988	0.1182	73
1721	Multiynthesizer (Idefics2-8B)	0.2918	0.2952	0.4684	0.4741	0.1152	0.1061	75
1723	Multiynthesizer (LlaMa)	0.2710	0.2946	0.4548	0.4747	0.0873	0.1145	76
1957	Synthesizer (Idefics2-8B) + MedCLIP Reranker	0.2852	0.2926	0.4728	0.4671	0.0976	0.1181	78
1669	LM-Fuser	0.2780	0.2894	0.4617	0.4611	0.0943	0.1177	79
1954	LM-Fuser + MedCLIP Reranker	0.2766	0.2874	0.4611	0.4602	0.0942	0.1146	80
1960	Mixer	0.2959	0.2853	0.4162	0.4157	**0.1756**	**0.1548**	83
646	Mixer + DMMCS ($\alpha = 0.1$)	0.2632	0.2720	0.4226	0.4217	0.1039	0.1222	92

In API calls, we used a temperature of 0.2 and top-p (Nucleus) sampling [36] with $p = 0.95$ to ensure consistency and reduce response variability. The model was queried in vision mode with high-resolution PNG inputs. Each image was processed once with the full list of terms, and the output consisted of the same image with labeled bounding boxes drawn directly onto it. We experimented with multiple prompt variants, refining them based on empirical alignment between model outputs and expected visual regions. The final prompt template is included in the Appendix 5.

The final evaluation was carried out by a radiologist, who assessed both the captions and their

Table 9
Detailed metric breakdown for our best-performing models on the test set.

Run ID	Overall	Similarity	BERTScore	ROUGE-1	BLEURT	Rel. Avg.	UMLS F1	AlignScore	Fact. Avg.	Rank
1403	0.**3068**	**0.7947**	0.5884	0.2176	0.3030	0.4759	**0.1429**	**0.1325**	**0.1377**	**48**
1463	0.3047	0.7942	0.5930	**0.2192**	0.3013	0.**4769**	0.1419	0.1230	0.1324	50
1462	0.3046	0.7940	0.5930	0.2191	0.3011	0.4768	0.1419	0.1230	0.1325	51
1717	0.3039	0.7939	0.5908	0.2174	0.3004	0.4757	0.1428	0.1213	0.1321	56
1968	0.3030	0.7886	**0.5950**	0.2150	0.2926	0.4728	0.1416	0.1250	0.1333	58
1724	0.3026	0.7896	0.5939	0.2122	0.2897	0.4714	0.1421	0.1257	0.1339	61
1718	0.2957	0.7844	0.5896	0.2148	0.3044	0.4733	0.1332	0.1031	0.1182	73
1721	0.2952	0.7898	0.5814	0.2138	0.3113	0.4741	0.1265	0.1061	0.1163	75
1723	0.2946	0.7917	0.5777	0.2171	**0.3121**	0.4747	0.1318	0.0972	0.1145	76
1958	0.2926	0.7725	0.5872	0.2082	0.3010	0.4672	0.1313	0.1048	0.1181	77

associated visualizations across multiple criteria, using a 5-point Likert scale (with 5 being the best score). Our submission ranked 1st in the Explainability task, out of a total of two participating teams. Table 10 presents the evaluation scores achieved by our system. Our strongest scores were in caption readability (4.5) and methodology appropriateness (4.0), reflecting the fluency and clarity of our outputs as well as the structured nature of our prompting approach. However, lower ratings were assigned for clinical appropriateness (2.7) and level of detail (2.6), indicating that while our captions were readable, they often lacked sufficient clinical specificity and depth. Similarly, the visualization focus score of 2.6 suggests that bounding boxes did not consistently align with the most salient medical regions. These results highlight both the promise and current limitations of prompt-based, supervision-free explainability systems in clinical imaging tasks.

Table 10
Evaluation results of our submission to the ImageCLEFmedical 2025 Explainability task. The system ranked 1st out of 2 teams based on human evaluation.

Metric	Score
Caption readability	4.5
Clinical appropriateness	2.7
Level of detail	2.6
Caption focus	3.3
Mean caption rating	3.3
Text coherence	3.1
Completeness	2.8
Visualization focus	2.6
Visualization rating	2.8
Methodology	4.0
Overall score	3.2
Rank	**1**

4.5. Hardware Configuration

For GPU acceleration, we used 1 NVIDIA Quadro 6000 GPU with 24GB memory for the training of each model.

5. Conclusions

Our participation in the ImageCLEFmedical Caption task provided an opportunity to explore innovative approaches that combine vision and NLP techniques for medical image captioning. Utilizing state-of-the-art models, we demonstrated competitive performance in Concept Detection, Caption Prediction, and Explainability tasks.

In the Concept Detection task, we achieved the 1st place out of 9 participating groups. Our top-performing system was an ensemble of CNN-FFNN models, combining multiple instances trained with different configurations (§3.1.3). Each individual model followed the CNN-FFNN pipeline (described in §3.1.1). We also applied a per-label thresholding strategy (§3.1.2) during tuning, which adjusted the decision threshold for each concept individually to optimize the F_1 score.

In the Caption Prediction task, our team ranked 5th out of 8 participating groups. Building on our previous work [16, 17, 37] and leveraging recent advancements in NLP—particularly instruction-tuned Large Language Models—we designed a multi-stage captioning pipeline. Our approach starts with the generation of initial captions using the InstructBLIP model [5]. These captions are then subsequently refined by incorporating information synthesized from captions of semantically similar images [9, 38], and further enhanced using a language model pre-trained on medical text [39] to improve clinical relevance and fluency.

In the Explainability task, our submission achieved the 1st place among the 2 participating teams. Our approach involved generating visual explanations that align with caption outputs by associating extracted medical entities with spatially localized regions in the radiology images. While our current method relies on an external model (GPT-4o) rather than the black-box captioning model itself, future work will explore more integrated explainability strategies—such as analyzing attention weights or saliency maps from the captioning model, enabling explanations that reflect its internal reasoning. This direction offers potential for more coherent and model-intrinsic justifications of predicted captions.

In future work, we plan to improve image preprocessing pipelines—particularly resolution normalization and modality-specific transformations—to enhance model robustness, following similar considerations raised by previous participants in the task [40]. We also intend to fully train our Mixer framework, allowing the reinforcement signal to play a more significant role throughout the training process. Additionally, we aim to develop a unified, multitask model capable of jointly addressing both the Concept Detection and Caption Prediction tasks. This joint framework will incorporate reinforcement learning signals from both tagging and captioning metrics, enabling more coherent and mutually informed outputs. Ultimately, we envision such systems as stepping stones toward clinically useful and trustworthy multimodal AI tools in radiology and beyond.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o exclusively as a component of the proposed system architecture for the explainability task, as thoroughly described in the respective sections (§3.3, 4.4, 5). The model was used via the OpenAI API)[7], to prevent any risk of competition data leakage. The authors did not use Generative AI tools, such as chatbots, for text creation, text translation, sentence polishing, image creation or rephrasing.

References

[1] 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,

[7] https://openai.com/api/

M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Sc hwab, 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.

[2] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. Ben Abacha, A. Garc'ıa Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[3] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2497–2506. doi:10.1109/CVPR.2016.274.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A Survey of Large Language Models, 2023. doi:10.48550/arXiv.2303.18223. arXiv:2303.18223.

[5] W. Dai, J. Li, D. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 49250–49267. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

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

[7] M. Samprovalaki, Exploring Multimodal Large Language Models for Medical Image Captioning, Master's thesis, Athens University of Economics and Business, Department of Informatics, 2024.

[8] P. Kaliosis, J. Pavlopoulos, F. Charalampakos, G. Moschovis, I. Androutsopoulos, A Data-driven Guided Decoding Mechanism for Diagnostic Captioning, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7450–7466. URL: https://aclanthology.org/2024.findings-acl.444/. doi:10.18653/v1/2024.findings-acl.444.

[9] G. Vernikos, A. Brazinskas, J. Adamek, J. Mallinson, A. Severyn, E. Malmi, Small Language Models Improve Giants by Rewriting Their Outputs, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 2703–2718. URL: https://aclanthology.org/2024.eacl-long.165/.

[10] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling Instruction-Finetuned Language Models, Journal of Machine Learning Research 25 (2024) 1–53. URL: http://jmlr.org/papers/v25/23-0870.html.

[11] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive Learning from Unpaired Medical Images and Text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3876–3887. URL: https://aclanthology.org/2022.emnlp-main.256/. doi:10.18653/v1/2022.emnlp-main.256.

[12] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-Critical Sequence Training for Image

Captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1179–1195. doi:10.1109/CVPR.2017.131.
[13] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, volume 2380 of *CEUR Workshop Proceedings*, 2019.
[14] B. Karatzas, J. Pavlopoulos, V. Kougia, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, volume 2696 of *CEUR Workshop Proceedings*, 2020.
[15] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21-24, volume 2936 of *CEUR Workshop Proceedings*, 2021, pp. 1184–1200.
[16] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.or, Bologna, Italy, 2022, pp. 1355–1373.
[17] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.
[18] M. Samprovalaki, A. Chatzipapadopoulou, G. Moschovis, F. Charalampakos, P. Kaliosis, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group in ImageCLEF medical 2024 (highlighted talk), in: Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 2024.
[19] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating Biomedical Terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.
[20] A. Chatzipapadopoulou, Enhanced Biomedical Image Tagging, Bachelor's thesis, Athens University of Economics and Business, Department of Informatics, 2025. URL: http://nlp.cs.aueb.gr/theses/Bsc_Thesis_Chatzipapadopoulou.pdf.
[21] F. Radenović, G. Tolias, O. Chum, Fine-Tuning CNN Image Retrieval with No Human Annotation, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 1655–1668. doi:10.1109/TPAMI.2018.2846566.
[22] D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
[23] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, 2019, pp. 6105–6114.
[24] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.
[26] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned Language Models Are Zero-Shot Learners, International Conference on Learning Representations abs/2109.01652 (2021). doi:10.48550/arXiv.2109.01652.
[27] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.
[28] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented

Generation for Large Language Models: A Survey, 2024. doi:10.48550/arXiv.2312.10997. arXiv:2312.10997.

[29] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neural Information Processing Systems abs/2005.11401 (2020).

[30] 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. G. 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.

[31] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. Rush, D. Kiela, M. Cord, V. Sanh, OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 71683–71702. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/e2cfb719f58585f779d0a4f9f07bd618-Paper-Datasets_and_Benchmarks.pdf.

[32] H. Laurençon, L. Tronchon, M. Cord, V. Sanh, What matters when building vision-language models?, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 87874–87907. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/a03037317560b8c5f2fb4b6466d4c439-Paper-Conference.pdf.

[33] 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: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021. URL: https://arxiv.org/abs/2103.00020.

[34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, International Conference on Learning Representations abs/1904.09675 (2019).

[35] M. Neumann, D. King, I. Beltagy, W. Ammar, SciSpaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 319–327.

[36] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, International Conference on Learning Representations abs/1904.09751 (2019).

[37] G. Moschovis, E. Fransén, NeuralDynamicsLab at ImageCLEF Medical 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[38] Y. Li, X. Liang, Z. Hu, E. Xing, Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation, in: AAAI Conference on Artificial Intelligence, volume abs/1903.10122, 2019. doi:10.1609/aaai.v33i01.33016666.

[39] Q. Lu, D. Dou, T. Nguyen, ClinicalT5: A Generative Language Model for Clinical Text, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 5436–5443. doi:10.18653/v1/2022.findings-emnlp.398.

[40] Q. V. Nguyen, H. Q. Pham, D. Q. Tran, T. K.-B. Nguyen, N.-H. Nguyen-Dang, T. B. Nguyen-Tat, UIT-darkcow team at imageCLEFmedical caption 2024: Diagnostic captioning for radiology images efficiency with transformer models, in: CLEF, 2024, pp. 1695 – 1710.

Table 11
Performance of exploratory models evaluated on our held-out development (private test) set. These models were not submitted to the official test set.

Model	F1 (dev)	F1 (val)
CNN + FFN (baseline)	0.5872	**0.5891**
Fine-Tuned → Ultrasonography	**0.5891**	0.5774
Masking (1)	0.5868	0.5773
Masking (2)	0.5806	—

Appendix

Explainability Task Prompt

Below we present the prompt template that yielded the most reliable outputs for bounding box generation during the Explainability Task. This structured instruction was passed to GPT-4o along with the medical image, the generated caption, and the extracted list of medical terms.

Prompt Template:

You are a radiologist with expert-level understanding of diagnostic imaging. I will provide you with three inputs:

1. A **medical image** (such as an X-ray, CT, or MRI),
2. A **caption** describing key findings in the image, and
3. A **list of medical terms** extracted from the caption.

Your task is to **perform image grounding** for the medical terms in the caption. This means:

- For each medical term, **draw a bounding box** around the corresponding anatomical or pathological feature in the image where it is visible.
- Label each bounding box with the **medical term**, using the **same color** for both box and label to ensure clarity.
- If a term refers to an **imaging modality** (e.g., "X-ray", "MRI", "CT"), draw a **neutral gray bounding box** around the **entire image** and label it accordingly.
- Ensure that all boxes are **tight, accurate**, and drawn based on radiological expertise.
- If a feature is **not clearly visible** or **ambiguous**, indicate the approximate region with a **dotted or lighter box** and note that the feature is inferred.

The final output should be:

- Visually clear, with minimal overlapping when possible;
- Consistent in label formatting and color coding;
- Suitable for educational or clinical use.

I will now provide you with the image, caption, and medical terms list.