

DS4DH Group at ImageCLEFmedical Caption 2025

Notebook for the DS4DH Lab at CLEF 2025

Jiawei He¹, Sohrab Ferdowsi², Weibo Feng², Fernando Alves⁴, Alexandra Platon^{2,3} and Douglas Teodoro^{2,*}

¹Hunan City University, 518 Yingbin East Road, Yiyang, Hunan Province, China

²University of Geneva, Chemin des Mines 9, 1202 Geneva, Switzerland

³Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 4, 1205 Geneva, Switzerland

⁴Pontifícia Universidade Católica do Rio de Janeiro Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro

Abstract

This paper presents the DS4DH team's approaches for the ImageCLEFmedical Caption 2025 challenge, where we participated in two tasks Concept Detection and Caption Prediction. For Concept Detection, unlike the typical approach of multi-label classification, we posed the problem as image to sequence of tokens mapping, where the tokens consist of the CUI's, and a standard transformer maps the image embeddings into the sequence. Our approach achieved an F1-score of 52.3%, ranking top-6 among the best submission systems. For Caption Prediction, we developed multiple approaches, including fine-tuned InstructBLIP, traditional retrieval-augmented generation (RAG), and cluster-based RAG methods. Our best strategy, based on the InstructBLIP model, achieved the highest recall (BERTScore (Recall) of 60.7%) among all participants, being ranked top-2 according to the Overall challenge metric (33.6%). Our experiments reveal that RAG approaches did not outperform the baseline, exposing critical challenges in medical image captioning where noisy retrievals significantly weaken generation performance. Through validation experiments and case studies, we demonstrate that only highly accurate reference images prove helpful, as poor retrieval quality introduces noise that degrades caption generation.

Keywords

ImageCLEF, RAG, Image Embedding, Radiology

1. Introduction

ImageCLEF is an evaluation forum that has been benchmarking multimodal information retrieval technologies since 2003, providing access to large collections of multimodal data across various domains, including medical imaging, social media, and Internet applications [1]. The ImageCLEFmedical Caption 2025 task focuses on automatic interpretation and summarization of medical images, addressing the time-consuming challenge of generating descriptive captions that can approximate the mapping from visual information to condensed textual descriptions typically performed by highly trained medical experts [2].

Our team, DS4DH, participated in two subtasks of ImageCLEFmedical Caption 2025: *Concept Detection* and *Caption Prediction*, achieving 6th and 2nd place, respectively. Notably, for the Caption Prediction task, we obtained the highest recall score (BERTScore [3] Recall) among all participating teams. In this paper, we present our approaches for both tasks and provide detailed observations from our experiments.

For the *Caption Prediction* task, we aimed to construct an effective retrieval-augmented generation (RAG) [4] framework for medical image captioning. Although our RAG approaches did not surpass the baseline on the test set, we observed critical challenges in applying RAG to radiology report generation: noisy information and poor retrieval quality can significantly weaken the model's generation capability. We validated this observation through validation set experiments and provided case studies with specific

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

✉ joewellhe@gmail.com (J. He); sohrab.ferdowsi@unige.ch (S. Ferdowsi); Weibo.Feng@etu.unige.ch (W. Feng);

fernando1_ala@hotmail.com (F. Alves); alexandra.platon@hug.ch (A. Platon); Douglas.Teodoro@unige.ch (D. Teodoro)

🆔 0009-0005-3608-3869 (J. He); 0000-0003-3768-6408 (S. Ferdowsi); 0000-0002-2043-3052 (W. Feng); 0009-0009-6946-208X (F. Alves); 0000-0001-6238-4503 (D. Teodoro)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

examples. Our findings highlight the importance of filtering noise information and identifying more precise references to build effective RAG-based radiology report generation systems, which will be the focus of our future research.

The source code for all our methodology and experiments on both sub-tasks is presented in this GitHub repository: <https://github.com/ds4dh/image-clef-2025-med-radiology>.

2. Dataset

The Competition utilizes the dataset provided by ImageCLEFmedical Caption Task 2025 [2], which employs the Radiology Objects in CONtext Version 2 (ROCOv2) dataset [5], an updated and extended version of the original ROCO dataset [6]. The dataset comprises curated medical images from biomedical literature in the PMC OpenAccess subset, accompanied by their corresponding captions and manually controlled UMLS [7] terms as metadata.

The dataset contains a total of 116,635 radiology images distributed across three splits: 80,091 images for training, 17,277 for validation, and 19,267 for testing. The test set consists of previously unseen images to ensure robust evaluation. For the concept detection component, concepts are derived from a filtered subset of the UMLS 2022 AB release. The filtering process removes low-frequency concepts and restricts concepts based on their semantic types to improve recognition feasibility, following recommendations from previous challenge editions. For the caption, all captions undergo preprocessing steps including the removal of hyperlinks to ensure consistency across the dataset.

The ImageCLEFmedical Caption Task 2025 consists of two subtasks using the same medical image dataset. The concept detection task requires predicting a list of relevant CUI codes from the UMLS vocabulary that represent medical concepts present in the image. The caption prediction task involves generating descriptive captions for medical images that accurately describe the visual content and medical findings. Each image in the training and validation sets is annotated with both ground truth CUI codes and reference captions, while the ground truth codes and captions are not given in the test set.

3. Evaluation Metrics

In this section, we introduce the evaluation metrics used in the ImageCLEFmedical Caption Task 2025. The evaluation framework encompasses both concept detection accuracy and caption generation quality assessment.

3.1. Concept Detection

The concept detection task is evaluated using the F1-score calculated between predicted and ground truth CUI codes. The evaluation employs Python scikit-learn's F1-scoring method with binary averaging. For each image, binary arrays indicate the presence or absence of concepts in both predicted and ground truth sets. Two scores are reported: a primary score considering all concepts, and a secondary score filtering to manually annotated concepts only. The final score is averaged across all test images.

3.2. Caption Prediction

Caption evaluation is based on an average score across six metrics covering relevance and factuality aspects. All captions are preprocessed by converting to lowercase, replacing numbers with tokens, and removing punctuation.

Relevance Metrics:

- **Image-Caption Similarity:** Computed using medical imaging embeddings to measure similarity between image and caption representations

- **BERTScore** (Recall) [3]: Recall-oriented metric with IDF weighting using the microsoft/deberta-xlarge-mnli model for contextualized embeddings
- **ROUGE-1** [8]: F-measure score evaluating unigram overlap between generated and reference captions
- **BLEURT** [9]: Learning-based metric using BLEURT-20 checkpoint to assess text generation quality based on human judgments

Factuality Metrics:

- **UMLS Concept F1**: Evaluates medical concept accuracy using MedCAT for entity extraction and matching semantic types from the MEDCON [10] framework
- **AlignScore** [11]: RoBERTa-based metric assessing factual consistency by measuring information alignment between generated and reference texts

4. Methodology

In this section, we introduce the methods used by the DS4DH group for the two subtasks. Section 4.1 presents our approach for the concept detection task, while Section 4.2 describes our method for the caption prediction task.

4.1. Concept Detection

Solutions to the *concept detection* task consist typically of posing the problem as a multi-label classification task [12]. In our submissions, however, we consider the concept codes associated to each image as a sequence of codes and pose the problem as an image to sequence mapping task. This has two advantages. Firstly, one could use the powerful transformer-based sequence modeling [13] approaches that are prevalent in a wide range of tasks in modern machine learning. Secondly, since the concept codes (i.e., the UMLS CUI's) have typically a meaningful order into them, a sequence modeling task can capture this order. For example, the first codes typically contain modality tags (e.g., MRI, X-Ray, etc), while the following codes are anatomy-specific (abdomen, pelvis, upper), or disease-specific (Pleural effusion, metastatic malignant neoplasm to the liver, etc). Thanks to the position embedding of transformers, this order can be learned from the data and obviate the need for hand-crafted approaches trying to separate modality codes from other codes to help with the training, e.g., as in Figure 4 of [14].

In our architecture, we considered a convolutional neural network (CNN) [15] to embed images into a few low-dimensional vectors using heavy down-sampling. These vectors would correspond to the output channels of the CNN. A very small transformer decoder structure (with one single head and 2 layers) was used to attend to the image embedding vectors using cross-attention. The output sequence would be created after a linear layer mapping the transformer output to the vocabulary size of 2483 (2479 CUI's, plus 4 special tokens). During the model design phase, we noticed that the dimensions of the embedding layers, as well as the image embedding dimension could be reduced to as low as 16, without reducing performance, and also helping with the train-validation loss gap. This would give a network that has very small number of parameters (< 1 Mbytes).

As for the inference, the image embeddings corresponding to a new image would be fed to the transformer, while the beginning-of-sequence token would generate the output sequence autoregressively. Since the goal here is not to generate sequences with randomness, a beam-search was used to decode the outputs. We noticed that a beam search with size 3, would noticeably improve performance in comparison to a fully greedy decoding.

As has been pointed out by participants from previous challenges, the database has a large imbalance in terms of the frequency of the codes, as certain codes are much more present than others. In order to account for this, apart from the standard cross-entropy loss, we also tested with the focal loss [16] to encourage less common CUI's to be weighted more than easy examples. As an alternative strategy to improve the diversity of the predicted codes, we also tested with the idea of label-smoothing, i.e., to discount the probability of ground-truth token by a small fraction.

Our main submission does not use any external pre-trained resources, as all elements are exclusively trained on the provided training set of the challenge. This includes both the CNN and the transformer head. However, we also tested with pre-trained UMLS CUI weights from MedCPT [17], as well as CUI2Vec [18] pre-trained weights. In order to match the dimensionality, we used random projections to down-project the vectors to the dimension 16 and re-normalized. The embedding table of the CUI tokens were then replaced with these vectors, instead of the randomly initialized ones.

4.2. Caption Prediction

For the *caption prediction* task, we developed three distinct approaches to address the challenge of automatic medical image captioning. Our methodology encompasses i) vision-language models (VLMs) that directly learn multimodal representations between images and text; ii) RAG-based methods that retrieve related radiology images from the training set and leverage their retrieved captions to enhance caption generation; and iii) cluster-based models that utilize CUI code grouping strategies to improve retrieval quality.

4.2.1. Vision-Language Models

Following the approach of Panagiotis *et al.*[14] in ImageCLEFmedical Caption 2023, we employed InstructBLIP-Flan-T5-XL [19] as our base VLM for caption generation. InstructBLIP combines a pre-trained vision encoder with a large language model through a Q-Former architecture, enabling instruction-following capabilities for multimodal understanding.

Model configuration: To optimize training efficiency while preserving performance, we implement selective parameter freezing. Specifically, we freeze the entire vision encoder and the language model encoder. In the language model decoder, only the parameters beyond the first 300 (in parameter enumeration order) are fine-tuned, while the earlier ones remain frozen. Likewise, in the Q-Former encoder, only the parameters after index 150 are updated during training. This configuration results in 16.5% of parameters being trainable, focusing adaptation on the cross-modal alignment components.

Training strategy: The model is fine-tuned for 15 epochs with a learning rate of 5e-6 using Adam optimizer. We employ gradient clipping with a maximum norm of 1.0 to ensure training stability. The instruction prompt guides the model to act as an experienced radiologist, generating descriptive captions that highlight location, nature, and severity of abnormalities in radiology images. To optimize memory usage, we load model parameters in bfloat16 precision, enabling training completion on a single Tesla V100 (32GB) GPU.

Inference configuration: During inference, we use beam search with 3 beams, maximum length of 80 tokens, and minimum length of 5 tokens. We apply a repetition penalty of 1.5 and prevent 3-gram repetitions to improve caption quality and diversity. Early stopping is implemented with a patience of 3 epochs based on validation loss to prevent overfitting.

4.2.2. Multi-Modal Retrieval-Augmented Generation

We implement an RAG framework that enhances caption generation by leveraging relevant training examples. This approach aims to address the challenge of generating contextually appropriate medical captions by incorporating knowledge from similar radiological cases.

Feature extraction and indexing: We utilize InstructBLIP's vision encoder to extract 1408-dimensional image features from the training set. Features are extracted from the pooled output (CLS token) of the vision model and normalized using L2 normalization to enable cosine similarity computation. A FAISS IndexFlatIP index is constructed to enable efficient similarity search across the 80,091 training images.

Retrieval strategy: For each test image, we extract visual features using the same encoder and retrieve the top-k most similar training images based on cosine similarity. We apply a similarity threshold of 0.95 to ensure only highly relevant cases are included in the context. The retrieved examples provide domain-specific knowledge that guides the generation process.

Context integration: Retrieved captions are incorporated into an enhanced instruction prompt that provides the model with relevant contextual examples. The RAG-enhanced prompt follows the format: "Based on these similar cases and what you see in the current image, generate a detailed caption," where similar cases are presented as reference examples before the generation task.

Generation process: The enhanced prompt containing both the base instruction and retrieved examples is processed through the fine-tuned InstructBLIP model. Those samples without any highly relevant cases will be generated only by the base instruction.

4.2.3. Cluster-based RAG

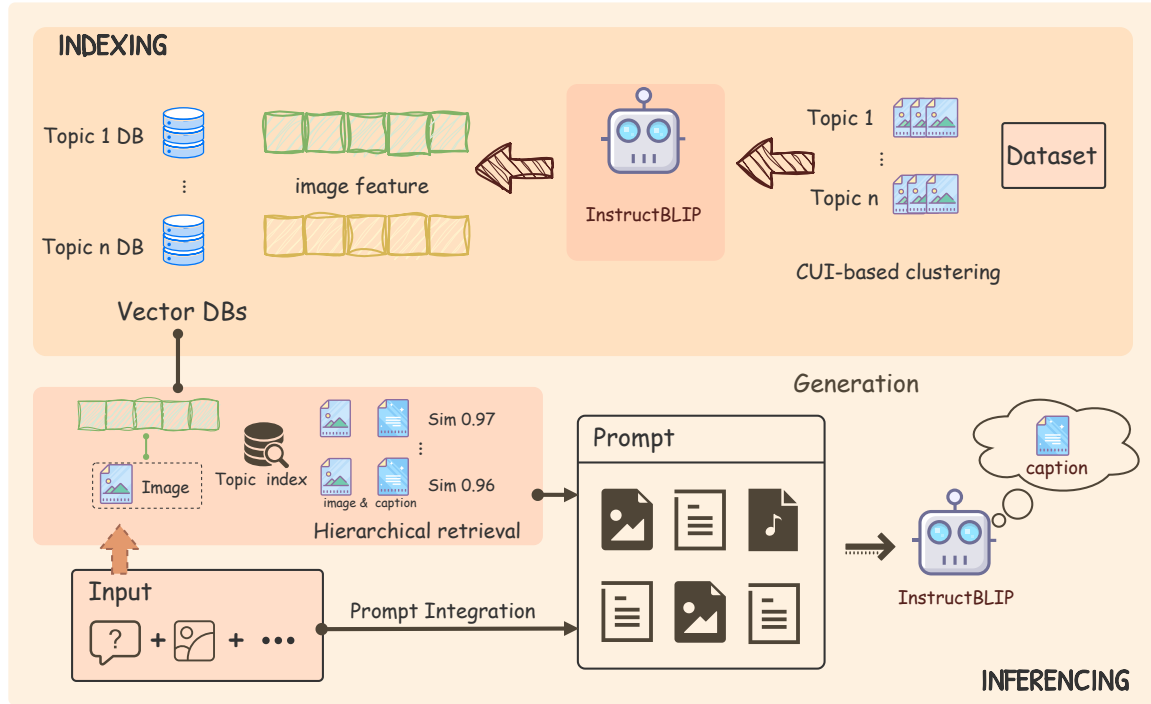


Figure 1: The workflow of the cluster-based RAG

To improve retrieval accuracy, we propose a cluster-based RAG approach that leverages medical concept similarity for enhanced context selection. This method aims to address the challenge of retrieving examples with medical relevance by organizing the training data into topic-specific clusters. Figure 1 shows the complete workflow of our cluster-based RAG approach, illustrating how training data is clustered based on medical concept similarity and how relevant examples are retrieved for caption generation.

CUI-based clustering: We perform clustering based on CUI (Concept Unique Identifier) codes using medical concept embeddings generated by MedCPT-Query-Encoder [17]. Each CUI code is embedded using its name and NCI definition, creating 768-dimensional representations that capture semantic relationships between medical concepts. We apply hierarchical clustering with a distance threshold criterion to automatically determine the optimal number of clusters based on the maximum allowable distance between cluster members. Training images are clustered based on the distance of their CUI embeddings, resulting in topic-specific groups that share similar medical concepts.

Topic-specific indexing: For each identified cluster, we build separate FAISS indices using InstructBLIP’s vision encoder features. This creates a two-level retrieval system where images are first categorized by medical topic, then retrieved based on visual similarity within relevant topics. A fallback index handles images without CUI annotations or those belonging to unrecognized topics.

Hierarchical retrieval strategy: During inference, we first identify relevant topic clusters based

on the test image’s CUI codes, then perform similarity search within those specific indices. We apply Reciprocal Rank Fusion (RRF) to merge results from multiple topic indices, weighting each result by both visual similarity and CUI semantic similarity. Similarity thresholds of 0.9 and 0.95 are applied to filter retrieved examples based on the cosine similarity of CUI embeddings and visual embeddings.

Context-aware generation: Retrieved examples are ranked by their combined RRF score and CUI similarity, ensuring that the most semantically and visually relevant cases guide the generation process. Similarly, those samples without any highly relevant cases will be generated only by the base instruction. For the validation set, we use ground-truth CUI codes for our hierarchical retrieval, while for the test set, we use our predicted results of Concept Detection task.

4.2.4. Alignment Model

Also motivated by Panagiotis *et al.* [14], we employ an additional language model to improve the generated captions through an extra alignment process.

Training data generation: We first use our fine-tuned InstructBLIP model to generate captions for the entire training set, creating input-output pairs where the generated captions serve as inputs and the ground truth captions as targets. This process produces a specialized dataset of 80,091 caption pairs that captures the specific improvement patterns needed for medical image captioning refinement.

Alignment model training: We utilize BioBart-v2-large [20] as our alignment model, training it to transform initially generated captions into more accurate descriptions. The model is trained for 20 epochs with a learning rate of $3e-5$ using the instruction prompt that guides the model to act as a medical professional enhancing generated captions. During inference, beam search with 3 beams and repetition penalty of 1.5 ensures diverse and high-quality caption refinement.

5. Results and Discussions

5.1. Concept Detection

Our concept detection submission ranked as the 5th team on the official leaderboard of the 2025 challenge. Below we elaborate on the results followed by a short discussion.

5.1.1. Submission Results

Table 1 sketches the results of our different models on the validation set of the challenge. We submitted only the baseline model to the challenge, as other ideas did not provide better results.

Table 1

Performance comparison of methods for image-to-CUI sequence prediction.

Method	F1-score (valid. set)	F1-score (test. set)	CUI length (avg.)	Total unique CUIs predicted
Image channel embedding to CUI seq. (our submitted baseline)	0.528	0.523	1.3	15
+ Focal loss ($\lambda = 1.5$)	0.502	-	1.3	15
+ Label smoothing	0.514	-	1.3	15
+ MEdCPT pretrained CUI emb.	0.527	-	1.3	15
+Short-length masking ($l = 3$)	0.393	-	3.0	103

5.1.2. Discussion

Our best result was achieved by our baseline submission. This means that the application of the focal-loss, the idea of label-smoothing, or the use of pre-trained embedding weights for the CUI’s did not bring meaningful improvements to the F1-score.

While this may seem counterintuitive, we hypothesize that the reason relates to a fundamental shortcoming of the database and its evaluation, namely, the very large imbalance of the length of the ground-truth codes in the training set. As is seen in 2, around 10 % of the training set consists of images for which only one single CUI, typically modality codes (MRI, CT, X-Ray, etc), has been attributed as the ground-truth.

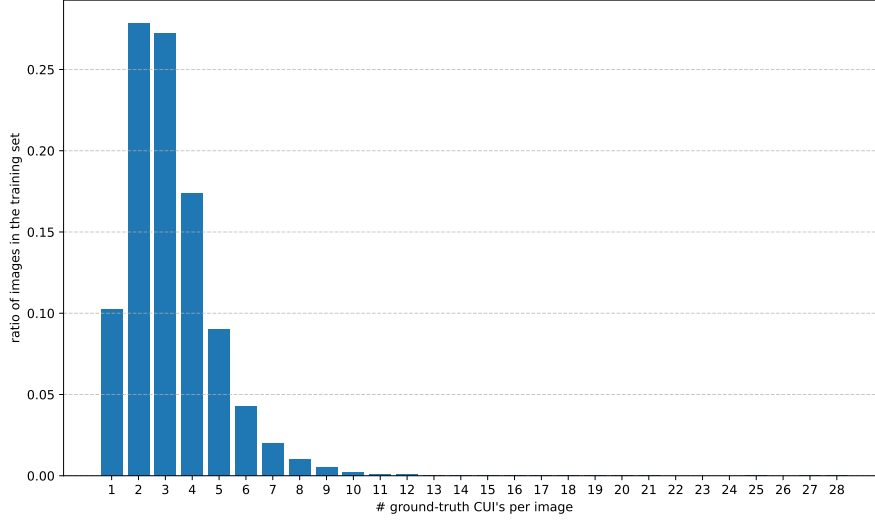


Figure 2: Histogram of the distribution of ground-truth code lengths in the training set.

This means that, during training, while the model is encouraged to predict more detailed codes from the examples with sufficient number of ground-truth CUI’s, at the same time, it is discouraged to learn more specific codes when encountering short CUI sequences (which happen to be at the majority).

Of course, this could be mitigated by adopting masking strategies during training, such that too short sequences with CUI lengths less than a threshold (say e.g., 3, as a reasonable expected length for a given medical image) are considered examples with missing codes, and are hence masked from the loss in a way that the model is not discouraged to predict longer CUI code sequences for them. However, as for the evaluation at the validation set or the test set of the challenge, this still is problematic, since the F1-score still penalizes predicted sequences that are longer than the ground-truth. In other words, if the model correctly learns to assign 3 CUI’s to a given image (e.g., specifying the imaging modality, the anatomy part, as well as a disease-related code), it is still penalized by the F1-score, since its corresponding ground-truth only specifies the code related to the imaging modality.

Note that there is no way that short code sequences could be learned from the images themselves, since the short codes are only artifacts of the original database curation procedure, rather than image-related attributes. This is why, we always noticed a certain gap between the training and validation loss-values, such that the model would overfit to the training set in spite of very heavy regularization or adopting extremely small models to avoid overfitting, or even freezing out large portions of the network.

Therefore, we propose that the evaluation procedure for the concept detection sub-task to adapt to this phenomenon, e.g., by not penalizing solutions with longer CUI’s for the heavily under-assigned samples in terms of CUI’s (e.g., those with sequence lengths less than or equal to 4).

Notice that, as depicted in table 1, the diversity of the CUI’s predicted by these models are very low (15), and the predicted CUI sequence lengths are also very low (≈ 1.3).

As for the last line of this table, i.e., masking the loss during training not to penalize CUI sequences smaller than 3, we see that the diversity sharply increases to 103 unique CUI’s, and also the average length increases to 3. However, since the final evaluation metric is the plain F1-score that penalizes longer than ground-truth CUI sequences, the performance drops, even though we believe that the model is more useful.

5.2. Caption Prediction

Overall, our group ranked 2nd in the Caption Prediction task, achieving first place in recall-based metrics (BERTScore). Among all our proposed approaches, the fine-tuned InstructBLIP model achieved the best performance, which significantly contributed to our team’s final ranking.

5.2.1. Submission Results

As shown in Table 2, the fine-tuned InstructBLIP model, treated as our baseline, achieved the best overall performance with a score of 0.3708, particularly excelling in BERTScore (Recall) (0.6067), which contributed to our first-place ranking in recall-based metrics.

Table 2

Caption prediction performance comparison across different runs using the official test set. Bold scores indicate the best performance for each metric.

Method	Overall	Similarity	BERTScore	ROUGE	BLEURT	UMLS F1	AlignScore
InstructBLIP	0.3708	0.9016	0.6067	0.2516	0.3096	0.1682	0.1417
+Alignment	0.3564	0.8626	0.6010	0.2497	0.3064	0.1545	0.1286
RAG	0.3478	0.8672	0.5952	0.2296	0.3099	0.1519	0.1110
Cluster RAG (0.95)	0.3526	0.8740	0.5983	0.2339	0.3094	0.1554	0.1176
Cluster RAG (0.97)	0.3620	0.8902	0.6034	0.2420	0.3087	0.1608	0.1300
+DeepSeek	0.3286	0.8516	0.5814	0.2058	0.3082	0.1317	0.0900

We also notice that the alignment model did not achieve effective improvements compared to the baseline InstructBLIP. The RAG approach failed to meet our expectations, with all metrics showing decline compared to InstructBLIP. We attribute this to the fact that while radiology images are visually similar, they can differ significantly in medical concepts. Consequently, the RAG approach introduced noise information rather than helpful context, leading to performance degradation.

The cluster-based RAG partially mitigated this degradation but still failed to outperform InstructBLIP. Notably, when we increased the CUI embedding similarity threshold from 0.95 to 0.97, the model’s performance approached that of InstructBLIP more closely (overall score improving from 0.3526 to 0.362). However, since our CUI similarity calculation was based on predicted CUI codes, which achieved only a 0.5225 F1 score, substantial noise remained in the system. We also attempted to use reference captions retrieved by Cluster RAG (0.97) along with the generated captions as input, prompting DeepSeek R1 to improve them. However, this approach failed to enhance performance because DeepSeek R1 cannot process visual image information.

Overall, our runs represent a key challenge we learned from applying RAG to radiology image caption generation in ImageCLEFmedical Caption 2025: only highly accurate reference images prove helpful, as inaccurate retrievals introduce noise that weakens the generation model’s performance.

5.2.2. Discussion

The validation set results presented in Table 3 are consistent with our test set findings, providing valuable insights into the effectiveness of different retrieval-augmented generation approaches for radiology image captioning.

Both the Alignment model and traditional RAG consistently underperform compared to the baseline InstructBLIP model across all evaluation metrics. Specifically, the Alignment approach shows degraded performance with BERTScore dropping from 0.6065 to 0.5955, while traditional RAG exhibits even more larger drops, achieving only 0.5952 BERTScore and 0.2396 ROUGE scores. This degradation suggests that introducing additional information without careful filtering can introduce noise that harms the generation process.

Cluster RAG results varied significantly based on CUI code quality. With predicted CUI codes (F1=0.5281), performance remained below InstructBLIP levels across most metrics. Using ground truth

Table 3

Caption prediction performance comparison in the validation set. Bold scores indicate the best performance for each metric

Method	BERTScore	ROUGE	BLEURT	AlignScore
InstructBLIP	0.6065	0.2617	0.3088	0.1364
+Alignment	0.5955	0.2507	0.3035	0.1220
RAG	0.5952	0.2396	0.3096	0.1034
Cluster RAG (0.95)	0.6056	0.2600	0.3089	0.1304
+Ground Truth CUI	0.6075	0.2618	0.3148	0.1209

CUI codes, however, improved three of four metrics over the baseline: BERTScore increased to 0.6075, ROUGE to 0.2618, and BLEURT to 0.3148.

The contrast between predicted and ground truth CUI performance points to a fundamental limitation: while Cluster RAG works well with accurate medical concepts, current CUI prediction methods introduce too much noise. Inaccurate CUI codes lead to poor topic clustering and irrelevant image retrieval, which then degrades caption quality. The retrieval system selects less useful reference materials when concept identification fails.

These findings highlight a key challenge in applying RAG to specialized medical domains: the critical importance of high-quality concept extraction and similarity matching. While our approach demonstrates the potential for improvement when accurate medical concept identification is achieved, future work should focus on developing more robust CUI prediction methods or alternative concept identification strategies to bridge the gap between ground truth and predicted performance.

5.2.3. Case Study

To illustrate the clustering quality variation, we present two representative cases from our clustering results in Table 4. Cluster 144 demonstrates strong semantic coherence, with all concepts relating to vascular system structures and blood vessels, including various arteries, veins, and vessel-related anatomical structures such as carotid, renal, iliac, and femoral vessels. In contrast, Cluster 403 exemplifies poor clustering quality, containing heterogeneous concepts spanning unrelated medical domains such as dental caries, gynecological conditions, surgical procedures, and anatomical structures. This difference shows that our clustering method is still in its early stages, with significant potential for enhancing clustering quality.

Currently, we evaluate clustering quality through manual annotation by experienced medical practitioners on a randomly sampled subset of 5 clusters containing approximately 100 CUI codes. While this approach provides insights into cluster coherence, developing comprehensive methods for systematic clustering quality assessment and achieving consistently better clustering performance remains an important direction for future research. The challenge lies in establishing automated evaluation metrics that can effectively capture semantic coherence across diverse medical concepts without requiring extensive manual review.

Table 5 presents two contrasting examples that illustrate the critical dependence of RAG performance on retrieval quality in medical image captioning.

Success case (Brain MRI): The RAG system demonstrates clear superiority over the baseline model. While the baseline incorrectly reports "no abnormalities," RAG accurately identifies "hyperintensities in the corpus callosum," which aligns with the ground truth describing bilateral white matter hyperintensities. This success stems from high-quality retrieval: both reference documents achieve exceptional similarity scores (0.984 and 0.980) and directly relate to the target pathology. The retrieved captions provide relevant medical concepts about white matter hyperintensities and corpus callosum involvement, enabling accurate caption generation.

Failure case (Chest X-ray): Conversely, RAG performs worse than the baseline generation. The baseline correctly identifies the bilateral nature with "bilateral pleural effusion," while RAG oversimplifies

Table 4

Case study: Comparison of good vs. poor clustering results.

Cluster 144 (Good Case)	Cluster 403 (Poor Case)
Analysis: Cohesive vascular system cluster Medical Concepts: <ul style="list-style-type: none"> - Structure of femoral artery - Femoral vein - Structure of iliac artery - Structure of jugular vein - Structure of renal artery - Structure of renal vein - Saphenous Vein - Veins - Vessel Positions - Structure of radial artery - Common carotid artery - Large blood vessel structure - Arterial system - Right common carotid artery structure - Left common carotid artery structure - Structure of right renal artery - Structure of left renal artery - Structure of internal iliac artery - Structure of external iliac artery - Internal jugular vein structure - Common Femoral Artery - Superficial femoral artery - Structure of left renal vein - Saphenous vein graft - Common iliac artery structure - Venous system - Collateral branch of vessel - Occluded Semantic Coherence: High - all concepts relate to blood vessels, arteries, veins, and vascular system structures	Analysis: Heterogeneous mixed medical concepts Medical Concepts: <ul style="list-style-type: none"> - Sampling [surgical action] - Infiltration - Dental caries - Intratendinous Route of Administration - Polyhydramnios - Fallopian Tubes - Composition - Consistency - Placenta Previa - Endometrioma - Aspirate (specimen) - Biopsy - Nodule - Autologous (qualifier value) - Lower jaw region - Synovial bursa - Tumor tissue sample - Availability of - Intrauterine Devices - Hydrophilicity - Sclerosis - Penetration Semantic Coherence: Low - concepts span multiple unrelated medical domains (dental, gynecological, surgical procedures, anatomical structures)

the complex multi-system pathology to "left lower lobe pneumonia." Despite high similarity scores, the retrieved references prove problematic: the first document describes esophageal pathology irrelevant to pulmonary findings, and the second document, though relevant, focuses on a specific unilateral condition that fails to capture the bilateral and multi-organ nature of the actual pathology.

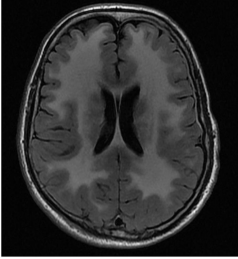
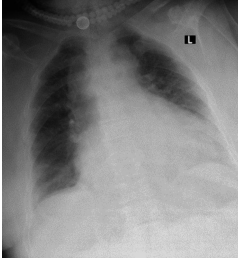
These contrasting cases demonstrate that similarity scores alone are insufficient indicators of retrieval quality. The success of RAG depends not only on semantic similarity but also on the medical relevance and comprehensiveness of retrieved references. When retrieval provides accurate, domain-specific medical concepts, RAG can correct baseline errors and improve caption quality. However, when retrieval introduces irrelevant information or overly specific references that miss the broader clinical picture, RAG performance degrades below baseline levels.

6. Conclusion

Our participation in ImageCLEFmedical Caption 2025 demonstrates both the potential and limitations of retrieval-augmented generation approaches for medical image captioning. While our fine-tuned InstructBLIP model achieved superior performance and the highest BERTScore among all participants,

Table 5

Case study: RAG performance analysis - good vs. bad cases.

Good Case: Brain MRI ImageCLEFmedical_Caption_2025_valid_12 CC-BY Katz (<i>et al.</i> 2023)	Bad Case: Chest X-ray ImageCLEFmedical_Caption_2025_valid_17 CC-BY Katagiri (<i>et al.</i> 2023)
Ground Truth Image: 	Ground Truth Image: 
Ground Truth Caption: MRI brain showing confluent, bilateral white matter T2 hyperintensity with involvement of the corpus callosum	Ground Truth Caption: Plain chest radiography showing bilateral hilar and lung congestion with scattered ground-glass opacities, obliteration of both costophrenic angles, enlarged cardiac shadow, and widened upper mediastinum
Generated (Baseline): Axial T1-weighted MRI of the brain showing no abnormalities . FLAIR: fluid-attenuated inversion recovery.	Generated (Baseline): Chest X-ray on admission showing bilateral pleural effusion.
Cluster RAG Result: Axial T2W MRI of the brain shows hyperintensities in the corpus callosum.	Cluster RAG Result: Chest X-ray showing left lower lobe pneumonia.
Retrieved References: Doc 1: White matter hyperintensities on T2-FLAIR sequence (Sim: 0.984, CUI: 0.941) Doc 2: Axial T2W MRI shows hyperintensity in corpus callosum (Sim: 0.980, CUI: 0.908)	Retrieved References: Doc 1: Anterio-posterior CXR views demonstrate dilated esophagus and air fluid level (Sim: 0.971, CUI: 0.938) Doc 2: AP CXR demonstrates left lower lobe pneumonia (Sim: 0.977, CUI: 0.910)
Analysis: RAG Success: - Corrected baseline's false negative - High-quality, semantically coherent retrieval - Accurate medical concept identification - Both references highly relevant to pathology Outcome: RAG > Generated	Analysis: RAG Failure: - Oversimplified complex multi-system pathology - Doc 1 irrelevant (esophageal vs pulmonary) - Doc 2 too specific, missed bilateral nature - Generated baseline was more accurate Outcome: Generated > RAG

our RAG experiments revealed fundamental challenges in applying retrieval methods to specialized medical domains. The critical finding from our work is that RAG performance heavily depends on retrieval quality: high-quality, medically relevant references can improve caption generation, but poor retrieval introduces noise that degrades performance below baseline levels. Our cluster-based approach partially addressed these issues, showing improved performance when using ground truth CUI codes compared to predicted ones, highlighting the importance of accurate medical concept identification. Future research should focus on developing more robust retrieval filtering mechanisms and improved medical concept extraction methods to bridge the gap between the theoretical potential and practical effectiveness of RAG systems in medical image captioning.

As for the concept detection sub-task, our novelty is at posing the problem as an image to sequence mapping, rather than the standard multi-label classification. Moreover, we highlight an important shortcoming with the evaluation protocol of the challenge that penalizes model predictions with CUI sequences longer than the available ground-truth. We suggest that changing the F1-score to a masked

F1-score would promote models trained with more diverse CUI's.

Declaration on Generative AI

The authors have not employed any Generative AI tools for writing this manuscript.

References

- [1] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ștefan, 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. Malvey, 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.
- [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. García 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] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [4] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meeting llms: Towards retrieval-augmented large language models, in: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, Association for Computing Machinery, New York, NY, USA, 2024, pp. 6491–6501. URL: <https://doi.org/10.1145/3637528.3671470>. doi:10.1145/3637528.3671470.
- [5] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco 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.
- [6] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in Context (ROCO): A Multimodal Image Dataset, Springer International Publishing, 2018, p. 180–189. doi:10.1007/978-3-030-01364-6_20.
- [7] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating Biomedical Terminology, *Nucleic Acids Research* 32 (2004) D267–D270. doi:10.1093/nar/gkh061.
- [8] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", in: *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013/>.
- [9] T. Sellam, D. Das, A. Parikh, "BLEURT: Learning Robust Metrics for Text Generation", in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. doi:10.18653/v1/2020.acl-main.704.
- [10] W. W. Yim, Y. Fu, A. Ben Abacha, et al., Aci-bench: a Novel Ambient Clinical Intelligence Dataset

- for Benchmarking Automatic Visit Note Generation, *Scientific Data* 10 (2023) 586. doi:10.1038/s41597-023-02487-3.
- [11] Y. Zha, Y. Yang, R. Li, Z. Hu, "AlignScore: Evaluating Factual Consistency with A Unified Alignment Function", in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. doi:10.18653/v1/2023.acl-long.634.
 - [12] J. Rückert, A. Ben Abacha, A. G. S. d. Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 - Caption Prediction and Concept Detection, in: *CLEF 2024 Conference and Labs of the Evaluation Forum*, 2024. URL: <https://www.microsoft.com/en-us/research/publication/overview-of-imageclefmedical-2024-caption-prediction-and-concept-detection/>.
 - [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, *Advances in neural information processing systems* 30 (2017). URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
 - [14] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2023, in: *CLEF (Working Notes)*, 2023, pp. 1524–1548. URL: <https://ceur-ws.org/Vol-3497/paper-126.pdf>.
 - [15] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, *IEEE Transactions on Neural Networks and Learning Systems* 33 (2022) 6999–7019. doi:10.1109/TNNLS.2021.3084827.
 - [16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020) 318–327. doi:10.1109/TPAMI.2018.2858826.
 - [17] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, MedCPT: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot Biomedical Information Retrieval, *Bioinformatics* 39 (2023) btad651. doi:10.1093/bioinformatics/btad651.
 - [18] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, I. S. Kohane, Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data, in: *Pacific Symposium on Biocomputing*. Pacific Symposium on Biocomputing, volume 25, 2020, p. 295. doi:10.1142/9789811215636_0027.
 - [19] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023. URL: <https://openreview.net/forum?id=vvoWPYqZJA>.
 - [20] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 97–109. doi:10.18653/v1/2022.bionlp-1.9.