

Retrieval Augmented Generation and Patch-Level Explainability applied to Medical Image Captioning with Small Visual Language Models

Notebook for the gfreitas team at CLEF 2026

Gabriel Freitas^{1,†}, Maria Bosco^{1,†}, Roberto Lotufo¹, Jayr Pereira^{1,2} and Diedre Carmo¹

¹ School of Electrical and Computer Engineering, Universidade Estadual de Campinas, Campinas, Brazil

² Centro de Ciências e Tecnologia, Universidade Federal do Cariri (UFCA), Juazeiro do Norte, Brazil

Abstract

This paper describes our participation in the ImageCLEF 2026 Medical Caption and Explainability tasks. For the captioning task, we investigate whether Retrieval-Augmented Generation (RAG) with similar cases improves Visual Language Model (VLM) based captioning quality compared to one-shot prompting strategies. We evaluate three prompting approaches — one-shot, random few-shot, and RAG-based few-shot using the top-3 most similar retrieved cases across three model configurations: MedGemma 4B, a LoRA fine-tuned variant of MedGemma 1.5 4B, and Qwen3-VL 4B. Case retrieval is performed using MedSigLIP embeddings and cosine similarity. For the explainability task, we propose an intrinsic, embedding-based interpretability method that segments the input image into spatial patches and the generated caption into semantic chunks, computing cosine similarity between their MedSigLIP embeddings to produce a heatmap that highlights the image regions most relevant to each part of the caption. Our system ranked 6th among 13 participating teams in the captioning task. Results suggest that RAG-based prompting improves caption quality, offering insights into the role of retrieved context in medical image captioning. Our team is registered as gfreitas on the ImageCLEF task platform.¹

Keywords

medical image captioning, retrieval-augmented generation, explainability, vision-language models

1. Introduction

Medical image captioning is a challenging task that requires integrating visual understanding with domain-specific clinical knowledge. Automatically generating accurate and informative descriptions for medical images has the potential to support radiologists and clinicians by reducing workload and improving documentation consistency. In this context, the ImageCLEF 2026 Medical Caption and Explainability [1, 2] tasks provide a benchmark for evaluating automated captioning systems and their interpretability across a diverse set of medical images from the Radiology Objects in COntext Version 2 (ROCOv2) dataset [3], an updated and extended version of the original ROCO dataset [4].

Large language models (LLMs) and vision-language models (VLMs) have shown remarkable capabilities in general image captioning. However, their application to the medical domain remains non-trivial due to the need for precise terminology, clinical reasoning, and domain-specific knowledge. In addition, real world deploy of VLMs have to consider ethical implications of using external, high footprint multimodal models through the public internet [5]. This has highlighted the need for performant "small" and local VLMs, able to run in modest consumer hardware. Here, we consider "small" VLMs those that fit in a single consumer graphics processing unit (GPU), generally with around 4B parameters. One such example in this direction is MedGemma [6], offering a promising direction by incorporating medical knowledge during training in a 4B model derived from Gemma [7]. Yet, generating high-quality captions still depends heavily on how the model is prompted and what contextual information is provided at inference time. Beyond caption quality, the lack of interpretability of these models is a critical barrier

¹Source code and experimental artifacts are publicly available at <https://github.com/unicamp-dl/rag-368>.

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

[†]These authors contributed equally to this work.

✉ g155421@dac.unicamp.br (G. Freitas); m183544@dac.unicamp.br (M. Bosco)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to their adoption in clinical settings, where understanding why a model produced a given output is as important as the output itself. Retrieval-Augmented Generation (RAG) [8] has emerged as an effective technique for grounding language model outputs in relevant external knowledge. By retrieving similar cases at inference time and incorporating them into the prompt, RAG can provide the model with concrete examples that guide caption generation without requiring additional fine-tuning. However, whether this approach yields consistent improvements over simpler prompting strategies in the medical captioning domain remains an open question.

In this work, we present our participation in both the ImageCLEF 2026 Medical Caption and Explainability tasks. For the captioning task, we systematically compare three prompting strategies: one-shot, random few-shot [9], and RAG-based few-shot using the top-3 most similar retrieved cases; across multiple model configurations, including pre-trained MedGemma 4B, MedGemma 1.5 4B [10] fine-tuned on ImageCLEF data, and pre-trained Qwen3-VL 4B [11]. For the explainability task, we propose an intrinsic, embedding-based interpretability method that segments the input image into spatial patches and the generated caption into semantic chunks, constructing a one-to-many heatmap that captures the alignment between generated caption chunks and image regions. Our results place us 6th among 13 participating teams in the captioning task, and provide insights into the relative contributions of retrieval-based context, model adaptation, and interpretability for medical image understanding.

2. Methodology

2.1. Dataset

Both the Medical Caption and Explainability tasks are based on the Radiology Objects in COntext Version 2 (ROCOv2) dataset [3], an updated and extended version of the original ROCO dataset [4]. ROCOv2 consists of curated radiology images sourced from biomedical articles of the PMC OpenAccess subset, accompanied by their original captions and associated UMLS terms provided as manually controlled metadata.

The dataset is split into three subsets: a training set of 97364 images, a validation set of 19240 images, and a test set comprising a previously unseen collection of images not available during development. The test set was provided by the challenge organizers exclusively for evaluation purposes.

2.2. Caption Prediction

We evaluate three model configurations for caption generation: MedGemma 4B (google/medgemma-4b-it) [10], a base vision-language model with medical domain pre-training; a QLoRA fine-tuned variant of MedGemma 1.5 4B (google/medgemma-1.5-4b-it), adapted by the authors on the ROCOv2 training set as described in Section 2.4; and Qwen3-VL 4B (Qwen/Qwen3-VL-4B-Instruct), a general-purpose multimodal model. All three are loaded in 4-bit quantized form and run with a maximum of 512 new tokens and a sampling temperature of 0.7. All experiments (inference and fine-tuning) were conducted at the Medical Image Computing Lab, School of Electrical and Computer Engineering, Universidade Estadual de Campinas (Unicamp), using a single Nvidia RTX 5090 GPU.

Each model is evaluated under three prompting strategies. The *one-shot* strategy sends the image with a single instruction: “Provide a precise and concise caption for this medical image.”. The *random few-shot* strategy injects three randomly selected captions from the training dataset before the query image, providing in-context demonstrations without any visual similarity criterion. The *RAG-based few-shot* strategy replaces random selection with similarity-based retrieval, injecting the captions of the three most visually similar training images into the prompt.

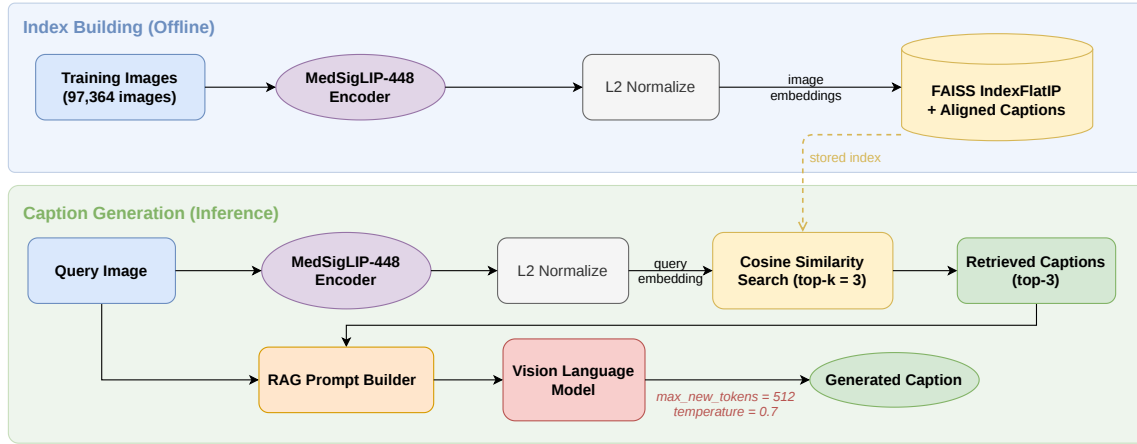


Figure 1: Overview of the RAG pipeline. Offline: MedSigLIP-448 embeddings are pre-computed for all 97,364 training images and indexed in a FAISS inner-product index alongside aligned captions. Inference: Each query image is encoded with the same encoder; the top-3 most similar images are retrieved and their respective captions injected into the prompt template before generation.

2.3. Retrieval Augmented Generation (RAG)

To enable similarity-based retrieval, we pre-compute visual embeddings for all 97,364 training images using MedSigLIP [10, 12] (google/medsiglip-448), a contrastive medical vision-language model trained on biomedical image-text pairs. For each image, the visual features are extracted from the `pooler_output` of the vision encoder and L2-normalized. The resulting embedding matrix is indexed offline using FAISS [13] with an inner-product index (IndexFlatIP), which is equivalent to cosine similarity over unit-norm vectors. The ground-truth captions are stored in a separate aligned list so that a retrieved index maps directly to its corresponding caption.

At inference time, each query image is encoded by the MedSigLIP encoder using the same normalization procedure. The resulting query vector is searched against the FAISS index to identify the top- k most visually similar training images. Initially, $k = 3$ is used in all submitted runs, and exploration of different retrieval parameters are left for future work. Their corresponding ground-truth captions are then passed to the generative model as context. Retrieved captions are injected into the prompt through a fixed template. The system prompt informs the model that it will receive captions of visually similar images, instructs it to use them only as medical context and style reference, and explicitly prohibits copying them. The user turn contains the query image followed by the retrieved captions formatted as a numbered list, closing with: *“Now provide a precise and concise caption for this medical image.”*

2.4. Fine-tuning

To adapt MedGemma to the medical captioning task, we employed QLoRA (Quantized Low-Rank Adaptation) [14, 15], a parameter-efficient fine-tuning technique that combines 4-bit weight quantization with low-rank adapter training, substantially reducing GPU memory requirements. The base model was loaded in 4-bit precision using the BitsAndBytes library, configured with the NormalFloat4 (NF4) quantization type, double quantization enabled, and BFloat16 as both the compute and storage data type. Double quantization further compresses memory usage by quantizing the quantization constants themselves. This configuration allowed fine-tuning of a 4B model on a single Nvidia RTX 5090 GPU.

Low-rank adapters were injected into all linear projection layers of the model (`target_modules = “all-linear”`), including the vision encoder and language model heads. The adapter rank was set to $r = 16$ with a scaling factor of $\alpha = 16$, and an adapter dropout of 0.05 was applied. The output projection layer (`lm_head`) and the token embedding layer (`embed_tokens`) were kept trainable in full precision via `modules_to_save`, ensuring the model could correctly map adapter outputs to vocabulary distributions.

Fine-tuning was performed using the SFTTrainer from the TRL library with the following hyperparameters: learning rate 2×10^{-4} , fused AdamW optimizer, linear learning rate schedule with a warmup ratio of 0.03, and maximum gradient norm of 0.3. Training ran for up to 5 epochs with a per-device batch size of 2 and 8 gradient accumulation steps, yielding an effective batch size of 16. Gradient checkpointing was enabled to reduce activation memory. The best checkpoint was selected based on minimum validation loss, with early stopping activated after a patience of 3 epochs with no improvement exceeding 10^{-7} , and at minimum 2 completed epochs before triggering.

Each training example was formatted using the model’s native chat template. The user turn contained the medical image together with the fixed instruction: “*You are a medical expert. Provide a precise and concise caption for this medical image.*” The ground-truth caption was placed in the assistant turn. Cross-entropy loss was computed exclusively over the assistant response tokens; image tokens and padding positions were masked with the ignore index (-100), preventing them from contributing to parameter updates.

RAG-integrated fine-tuning variant

A second fine-tuning variant (FT-RAG) augmented the user prompt with captions retrieved from the top-3 visually similar training images, identified via MedSigLIP embeddings and cosine similarity (Section 2.3). The retrieved captions were presented as medical context and style references within the user turn. This variant was trained with the same LoRA and quantization configuration as the caption-only model, but with 2 training epochs and evaluation every 500 steps. The label masking strategy remained identical: only the target caption in the assistant turn contributed to the training objective.

2.5. Explainability

To address the Explainability task, we propose an intrinsic, embedding-based interpretability method that establishes correspondences between regions of the input image and chunks of the generated caption, using multimodal embeddings generated by the model itself.

Given an input image, the model’s vision encoder extracts a sequence of patch embeddings arranged in a spatial grid of dimensions $H \times W$, where H and W are inferred from the model’s vision configuration. The hidden states of the last layer of the language model are extracted for the generated tokens, yielding one embedding vector per output token. Note this key design choice: by using the last-layer hidden states, comparisons are not isolated to the systematically aligned MedSigLIP module; instead, embeddings have gone through many transformer layers and are now distinct from the CLIP-based alignment strategy. Therefore, we hypothesize that high-dimensional alignment in this context is more meaningful than when isolated at the MedSigLIP level.

The generated token sequence is grouped into contiguous chunks to form semantically meaningful text zones. Three chunking strategies are supported by the provided code: (i) *fixed-size chunking*, which evenly divides the token sequence into a predefined number of chunks; (ii) *newline-based chunking*, which segments the text at paragraph boundaries; and (iii) *punctuation-based chunking*, which splits at punctuation marks (periods, commas, semicolons) and newlines. For each chunk, a single representative embedding is obtained by averaging the token embeddings within that span.

Both patch embeddings and chunk embeddings are L2-normalized. The relevance score of each patch p for a given chunk z is computed as the cosine similarity between their embeddings:

$$s(z, p) = \frac{\mathbf{e}_z \cdot \mathbf{e}_p}{\|\mathbf{e}_z\| \|\mathbf{e}_p\|} \quad (1)$$

This produces a similarity matrix of shape $[\text{Zones} \times P]$, where $P = H \times W$ is the total number of patches. A global similarity score is also computed by averaging all token embeddings into a single vector and computing its cosine similarity against all patches, yielding an overall saliency map for the full caption.

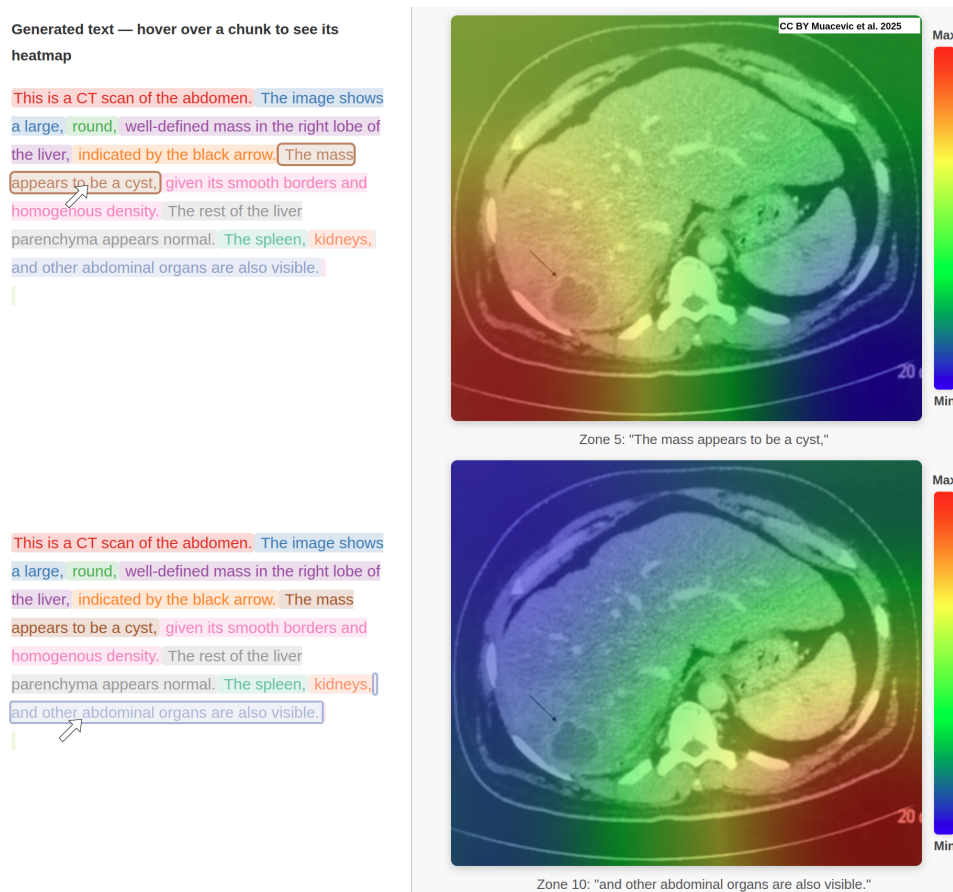


Figure 2: Example of interactive visualizations of the explainability method for pretrained MedGemma 1.5. In this case, using punctuation chunking, an average pool of 8 in patch embeddings, and a generic "write a paragraph caption" prompt. The arrow indicates cursor position.

The per-patch similarity scores for each chunk are mapped back onto the spatial grid and overlaid on the original image using a Blue-Green-Yellow-Red colormap, producing an interpretable heatmap. This allows inspection of which image regions are most relevant to each segment of the generated caption.

Beyond static heatmaps, our implementation supports the export of an interactive HTML report in which the generated caption is displayed alongside the medical image. As the user hovers over each caption chunk, the heatmap updates dynamically to highlight the image regions most associated with that specific text segment. This interactive interface provides a more intuitive and fine-grained tool for clinical inspection of model decisions, allowing users to explore the spatial grounding of each part of the caption without requiring any additional software (Fig 2). In the showcased example, the MedGemma 1.5 model without fine-tuning was prompted for a paragraph caption, outside of the style of ImageCLEF. Using punctuation chunking, each phrase is correlated with image regions. Notice how the explainability changes with different text chunks selected with the mouse cursor.

Submission configuration

In the final explainability task submission, we applied 8×8 average pooling to the patch embeddings prior to similarity computation (`pool_size=8`), reducing the spatial resolution of the patch grid and producing smoother, less noisy heatmaps. The caption was segmented using 4 tokens per chunk (`chunk_size=4`). This was necessary in an attempt to adapt to the shorter captions in the dataset.

3. Results and Discussion

We submitted our work to both the Caption Prediction task and the Explainability task under the *gfreitas* submission name. In the Caption Prediction task, we ranked 6th out of 13 submissions, and in the Explainability task, we ranked 3rd out of 3.

3.1. Caption Prediction Task

Table 1 summarizes seven of the most important configurations evaluated in the validation dataset. All metrics were computed using the provided evaluation scripts by the task organizers. Each run is identified by a short name used throughout this section and is characterized by its base model, prompting strategy, and whether any fine-tuning was applied. Other experiments are hidden from the notes, but can be found in the official repository.

Table 2 reports the validation set results for all configurations against the official metrics: Relevancy, Factuality, and their arithmetic mean (Overall), alongside BERT Score, ROUGE, semantic Similarity, BLEURT, MedCAT, and ALIGN. The task baseline results provided by the organizers is included for reference. Compared to the baseline, our submitted configurations underperformed on the validation set. Specifically, on Overall (0.343 vs. 0.331 best), Relevancy (0.534 vs. 0.519), BERTScore (0.609 vs. 0.600), ROUGE (0.271 vs. 0.228), and MedCAT (0.172 vs. 0.154). Our best experiments matched or exceeded the baseline on Similarity (0.932, tied), Factuality (0.151 vs. 0.164), and ALIGN (0.130 vs. 0.196).

Table 1

Description of evaluated configurations.

Run	Model	Prompt	Fine-tuning
MedGemma-OneShot	MedGemma 4B	OneShot	None
MedGemma-FewShot	MedGemma 4B	Few-shot	None
MedGemma-RAG	MedGemma 4B	RAG	None
MedGemma1.5-FT-Base	MedGemma 1.5 4B	OneShot	QLoRA on captions
MedGemma1.5-FT-RAG	MedGemma 1.5 4B	OneShot	QLoRA + RAG context
Qwen3-OneShot	Qwen3-VL 4B	OneShot	None
Qwen3-RAG	Qwen3-VL 4B	RAG	None

Across our submissions, Qwen3-RAG achieves the highest overall score (0.3315), followed closely by MedGemma1.5-FT-Base (0.3303) and MedGemma-RAG (0.3268). A consistent pattern across all models is the pronounced disparity between Relevancy and Factuality scores, as illustrated in Figure 3: Relevancy ranges from 0.457 to 0.519, while Factuality remains between 0.131 and 0.164. This gap indicates that the primary challenge lies not in producing topically relevant text, but in grounding generated content in clinically accurate facts.

Table 2

Validation set results. Baseline included for reference, bold marks the two best scores.

Run	overall	relevancy	factuality	similarity	bert	rouge	bleurt	medcat	align
Baseline	0.343	0.534	0.151	0.932	0.609	0.271	0.326	0.172	0.130
MedGemma-RAG	0.327	0.499	0.154	0.874	0.595	0.221	0.307	0.149	0.160
MedGemma-FewShot	0.305	0.478	0.131	0.836	0.582	0.205	0.290	0.113	0.150
MedGemma-OneShot	0.303	0.457	0.150	0.813	0.568	0.183	0.263	0.103	0.196
MedGemma1.5-FT-Base	0.330	0.512	0.149	0.925	0.588	0.228	0.307	0.154	0.143
MedGemma1.5-FT-RAG	0.319	0.498	0.140	0.884	0.587	0.221	0.301	0.145	0.135
Qwen3-RAG	0.331	0.519	0.144	0.932	0.600	0.217	0.327	0.153	0.135
Qwen3-OneShot	0.325	0.487	0.164	0.878	0.580	0.183	0.307	0.130	0.197

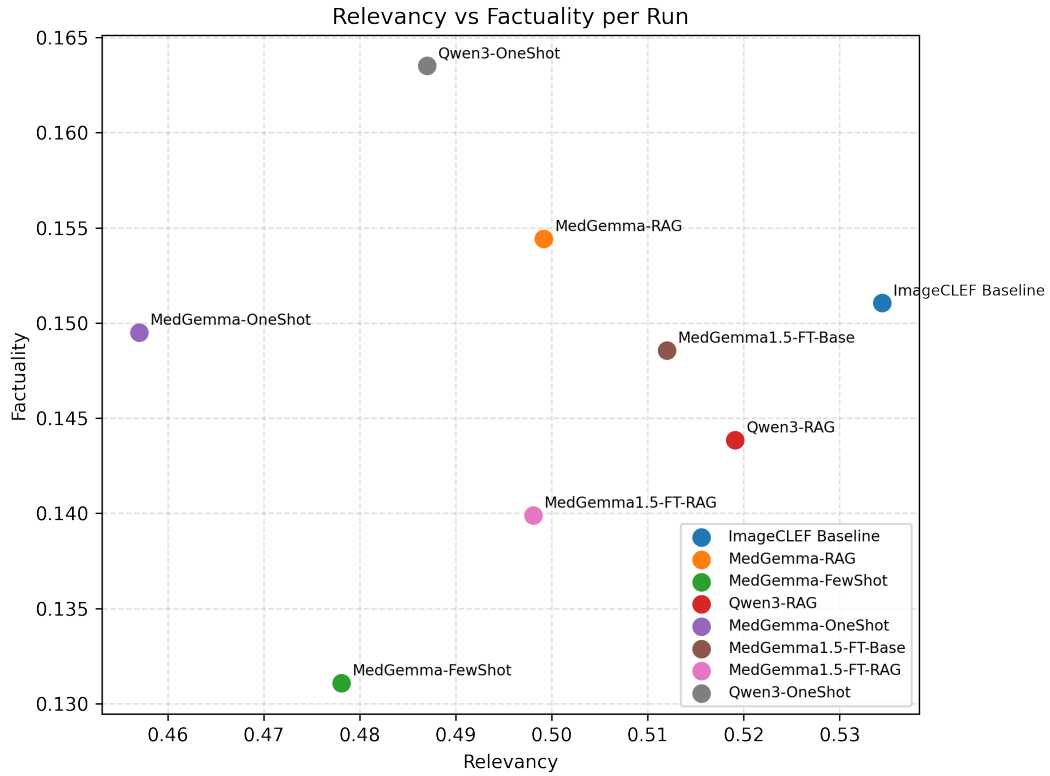


Figure 3: Relevancy and Factuality sub-scores for all our evaluated configurations on the validation set.

RAG-based prompting consistently outperforms simpler strategies within each model family. For MedGemma 4B, MedGemma-RAG achieves the highest overall score among the three MedGemma 4B variants (0.3268), above MedGemma-FewShot (0.3046) and MedGemma-OneShot (0.3033), suggesting that visually retrieved context provides more useful guidance than random demonstrations or no context at all. For Qwen3, Qwen3-RAG (0.3315) similarly outperforms Qwen3-OneShot (0.3253), though Qwen3-OneShot achieves the best factuality score across all runs (0.1635), indicating a trade-off between relevancy and factual precision depending on the prompting strategy.

Fine-tuning MedGemma 1.5 4B on the full ROCov2 training set (MedGemma1.5-FT-Base) yields the second-best overall score (0.3303), leading on ROUGE (0.2283) and MedCAT (0.1542), which reflect stronger lexical alignment and medical concept coverage. However, incorporating RAG context into the fine-tuning objective (MedGemma1.5-FT-RAG) reduces performance relative to the standard fine-tuned model (0.3190 vs. 0.3303), suggesting that retrieval is more beneficial when applied at inference time rather than during training.

Based on the validation results, the three best-performing configurations were selected for official submission to the test set. As shown in Table 3, all three maintain consistent performance on the test set, with slight overall score improvements relative to validation. Qwen3-RAG retains the top position (0.3343), followed by MedGemma1.5-FT-Base (0.3339), with a margin of less than 0.001 between them. MedGemma-RAG ranks third (0.3282) but achieves the highest factuality among the three test submissions (0.1608), closing the gap observed in validation.

An interesting result is that Qwen3-VL 4B, a general-purpose vision-language model with no specific medical pre-training, outperforms MedGemma 4B across all prompting strategies on both validation and test sets. This is somewhat unexpected, as MedGemma was explicitly trained on biomedical image-text data and might be expected to hold an advantage in a medical captioning task. One possible explanation is that Qwen3-VL’s stronger general vision-language capabilities and instruction-following behavior compensate for the lack of domain specialization, particularly when paired with RAG-retrieved captions that supply the necessary medical context at inference time. It is also possible that the evaluation

metrics reward fluent, well-structured descriptions that align more naturally with Qwen3’s generation style. That said, fine-tuning MedGemma 1.5 4B on the target dataset effectively closes this gap, with MedGemma1.5-FT-Base matching Qwen3-RAG within 0.001 overall on the test set. This suggests that medical domain pre-training remains valuable, but requires task-specific adaptation to be fully leveraged in this setting.

Table 3

Test set results.

Submission ID	Run	Overall	Relevancy	Factuality	Bert	Rouge	Similarity	Bleurt	MedCat	Align
840	Qwen3-RAG	0.3343	0.5201	0.1484	0.6013	0.2226	0.9273	0.3293	0.1563	0.1406
826	MedGemma1.5-FT-Base	0.3339	0.5147	0.1530	0.5887	0.2346	0.9254	0.3100	0.1602	0.1458
827	MedGemma-RAG	0.3282	0.4957	0.1608	0.5956	0.2210	0.8664	0.2999	0.1504	0.1711

3.2. Explainability Task and Qualitative Results

The final submission was evaluated by a specialist clinician across nine criteria, covering both the quality of the predicted captions and the adequacy of the accompanying visualizations. Scores were assigned on a Likert scale from 1 to 5, and an overall score was computed across all criteria. Our system achieved an overall score of 2.83.

Our best results were obtained on the caption quality criteria, whereas the visualization criteria revealed more limited performance, indicating room for improvement in our heatmap generation system.

One aspect we had already anticipated is that longer captions tend to produce heatmaps more closely aligned with the image content, as the chunking process generates more semantically relevant regions. However, to maintain consistency with the caption task, the final submission employed the ImageCLEF prompt, which resulted in shorter captions. We hypothesize that this domain shift likely constrained the quality of the resulting visualizations. Figure 4 illustrates qualitative results for four randomly selected validation cases under our test submission configuration, where these limitations can be seen. Note how Figure 2 shows the output for an alternative development prompt and configuration that produced a more detailed caption and, consequently, more expressive text chunks and a more accurate explainability heatmap.

A further limitation, closely related to the above, is the difficulty of finding a single optimal combination of chunking strategies and pooling configurations that generalizes well across all cases. Some images and captions benefited from a higher or lower configurations on chunk size (text) and pooling size (image). Nevertheless, we opted for a standardized configuration for the final submission. We will use these preliminary results to improve the explainability method in future work, aiming to achieve better generalization across different text sizes and styles.

4. Conclusion

This work explored the impact of RAG and Q-Lora fine-tuning for small visual language models performing medical image captioning in the context of the imageClef challenge. We compared three prompting strategies: one-shot, random few-shot, and RAG-based few shot across MedGemma 4B, a QLoRA fine-tuned variant of MedGemma 1.5 4B, and Qwen3-VL 4B, placing 6th out of 13 teams in the captioning task. RAG-based prompting consistently outperformed other strategies, and Qwen3-VL achieved better results than the medically pre-trained MedGemma across all prompting conditions, a gap that was effectively closed by fine-tuning MedGemma 1.5 on the target dataset. For the explainability task, we proposed an intrinsic explainability method that computes cosine similarity between vision encoding patch embeddings and last hidden layer output caption embeddings to produce spatial heatmaps, placing 3rd out of 3 teams. Our results are limited by the exploration only of 4B, relatively small models, and no comparisons to larger models to analyze possible efficiency gains from the

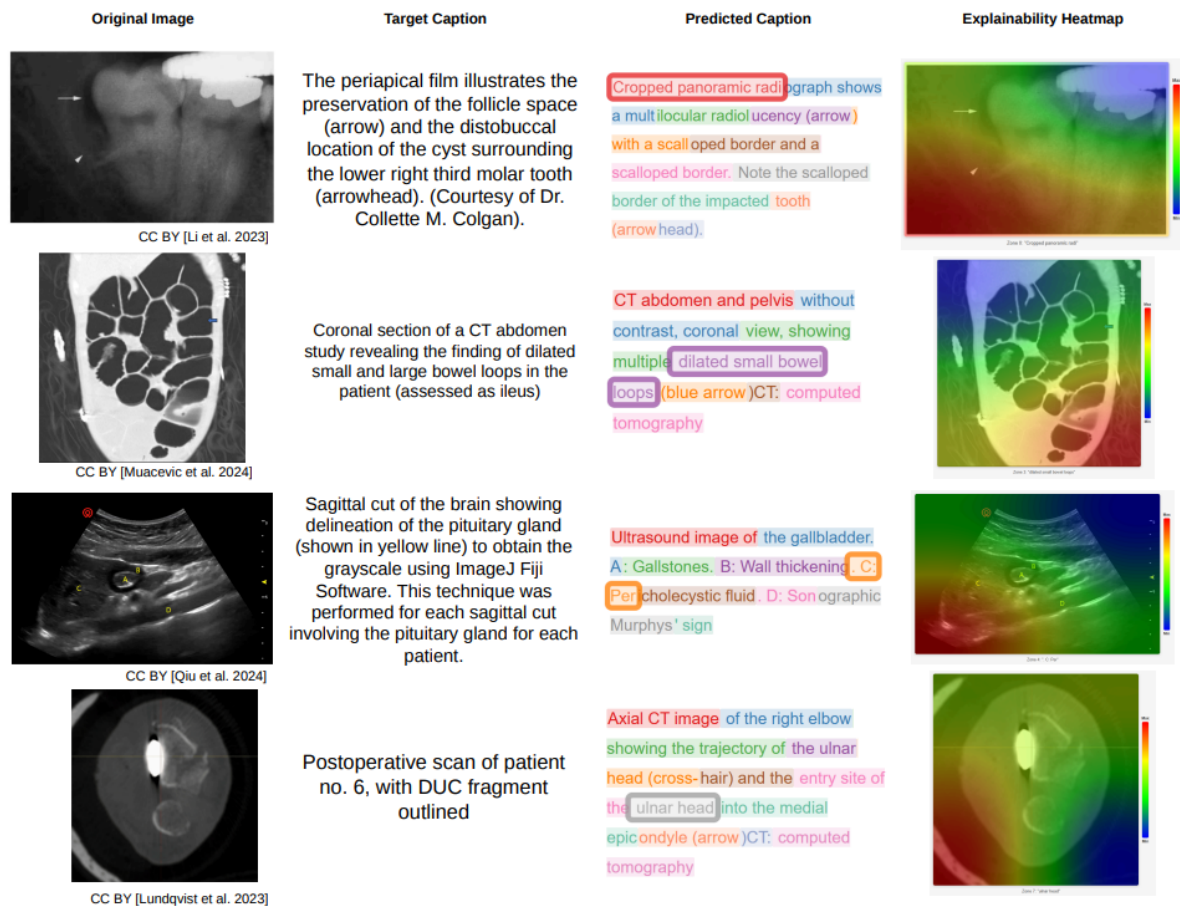


Figure 4: Example of a submitted result from both the Caption Prediction and the Explicability task. The predicted caption, generated using the same prompt adopted in section 2.3, produced a shorter description, resulting in broader and less precise heatmap regions. The highlighted parts of the captions, like "Cropped panoramic radi" is the chunk currently selected by mouse cursor position.

employed strategies. The explainability task solution underestimated the impact of small captions on the proposed method. Future work will explore multimodal RAG strategies (including generated text for retrieval queries), improvements to the explainability method, and applications to other datasets. We also did not analyze the quality of the retrieved cases and the retrieval failure-mode analysis is a priority for future work, alongside a closer look at which of the clinician's nine evaluation criteria most consistently degraded our explainability heatmaps.

The source code and all experimental artifacts are publicly available at <https://github.com/unicamp-dl/rag-368>.

5. Acknowledgements

This work was performed in part as a final project for the IA368 course at Universidade Estadual de Campinas (Unicamp). We thank the Medical Image Computing Lab for the hardware support.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4.5 in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [2] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryilmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org*, Jena, Germany, 2026.
- [3] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCov2: Radiology objects in context version 2, an updated multimodal image dataset, *Scientific Data* 11 (2024). doi:10.1038/s41597-024-03496-6.
- [4] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (ROCO): A multimodal image dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, Springer International Publishing, Cham, 2018, pp. 180–189.
- [5] D. S. Carmo, M. S. Sano, T. F. Moraes, L. Lopes, R. Dosso, G. V. Cruz, G. R. Garcia, J. R. Andrade, M. G. Saueressig, H. Mallmann, et al., Offline deployment of visual language models for automated prioritization of chest x-ray exams in a hospital environment, in: *Medical Imaging 2026: Imaging Informatics*, volume 13930, SPIE, 2026, pp. 138–143.
- [6] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al., MedGemma technical report, arXiv preprint arXiv:2507.05201 (2025).
- [7] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on Gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
- [8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: *NeurIPS*, 2020.
- [9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [10] A. Sellergren, C. Gao, F. Mahvar, T. Kohlberger, F. Jamil, M. Traverse, A. Tono, B. Sadjad, L. Yang, et al., MedGemma 1.5 technical report, arXiv preprint arXiv:2604.05081 (2026).
- [11] Q. Team, Qwen3 technical report, 2025. URL: <https://arxiv.org/abs/2505.09388>.

arXiv:2505.09388.

- [12] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: ICCV, 2023, pp. 11941 – 11952. doi:10.1109/iccv51070.2023.01100.
- [13] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Trans. Big Data 7 (2021) 535 – 547. doi:10.1109/tbdata.2019.2921572.
- [14] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: NeurIPS, 2023.
- [15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: ICLR, 2022.