

Model-Intrinsic Attention as Explanation for Radiology Image Captioning

Notebook for the CS_Morgan Lab at ImageCLEFmedical Caption 2026

Mahmudul Hoque¹, Raisa Nusrat Chowdhury¹, Ejiga Peter Ojonugwa Oluwafemi¹,
Okib Ul Islam¹, Rahmanul Hoque² and Md Mahmudur Rahman¹

¹Computer Science Department, Morgan State University, 1700 East Cold Spring Lane, Baltimore, Maryland 21251, USA

²School of Computer Information and Sciences, University of the Cumberlands 6178 College Station Drive, Williamsburg, Kentucky 40769, USA

Abstract

We describe CSMorgan's participation in the two ImageCLEFmedical 2026 subtasks, caption prediction and explainability, addressing both with a single family of vision-language models. LLaVA-OneVision-Qwen2, at the 0.5B and 7B scales, is adapted to radiology with two parameter-efficient fine-tuning methods, QLoRA and QDoRA, on the updated version of ROCov2 dataset, yielding four systems. Captions are produced by greedy decoding; explanations are produced by a model-intrinsic pipeline that reads the captioning model's own decoder attention over image patches, characterizes each layer as grounding-like or register-like, and renders attention heatmaps and bounding boxes on the input image. On the automated caption benchmark (15,249 test images; six metrics spanning relevance and factuality) our strongest system, 7B-QDoRA, placed fifth out of the participating teams inside a narrow top cluster. On the radiologist-judged explainability task the same system family placed first, winning six of nine criteria, and the radiologist identified our 7B-QDoRA captions as the most clinically accurate among the top 3 results. We examine the divergence between automated ranking and expert clinical judgment that participation in both subtasks makes visible.

Keywords

medical image captioning, vision-language models, explainability, attention, parameter-efficient fine-tuning, QLoRA, QDoRA, model-intrinsic explanation, visual grounding, ROCov2

1. Introduction

Interpreting and summarizing radiology images is a time-consuming task that requires highly trained experts and is a recurring bottleneck in clinical diagnosis pipelines. This has motivated sustained interest in automatic methods that map visual findings to concise textual descriptions, a problem that the ImageCLEFmedical caption task has tracked across successive editions on increasingly curated radiology data [1, 2]. Recent vision-language models (VLMs), which couple a visual encoder to a language decoder, have made fluent caption generation feasible, and parameter-efficient fine-tuning has made adapting such models to the radiology domain practical on modest compute budgets.

Fluent captions, however, are not on their own sufficient for clinical use. A caption does not indicate where in an image a model attended when it produced a given finding, and without that localization a clinician has no basis on which to trust or audit the output. Recognizing this, ImageCLEFmedical 2026 paired the caption-prediction subtask with an explainability subtask, asking participants to explain the captions for a small held-out set of images. The subtask placed no technical restrictions on the form of the explanation and was evaluated by a radiologist for interpretability, relevance, and creativity [1].

We enter both subtasks with a single family of systems. LLaVA-OneVision-Qwen2 [3] is fine-tuned on the updated version of ROCov2 [4] at two scales, 0.5B and 7B parameters, each with two parameter-

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

✉ mahoq1@morgan.edu (M. Hoque); racho1@morgan.edu (R. N. Chowdhury); ejiga.ojonugwa.peter@gmail.com (E. P. O. Oluwafemi); okisl1@morgan.edu (O. U. Islam); hoque.rahmanul@gmail.com (R. Hoque); md.rahman@morgan.edu (M. M. Rahman)

ORCID 0009-0006-5532-4135 (M. Hoque); 0000-0002-1663-6391 (R. N. Chowdhury); 0009-0003-2039-3075 (E. P. O. Oluwafemi); 0009-0000-8980-8273 (O. U. Islam); 0000-0003-0405-9088 (M. M. Rahman)



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

efficient methods, QLoRA [5] and QDoRA [6], giving four systems that share a base architecture, training data, and prompt. The same four systems serve the explainability subtask through a model-intrinsic pipeline: rather than annotating images with a separate post-hoc model, we extract the decoder self-attention of the captioning model itself over the image patches, so that the explanation is drawn from, and faithful to, the system that produced the caption. A central step in this pipeline characterizes each decoder layer’s attention as grounding-like or register-like, which lets us select the layers that carry input-dependent spatial signal rather than the fixed-position attention-sink patterns that a naive averaging of final-layer attention would surface.

Our results across the two subtasks are informative precisely in combination. On the automated caption benchmark (15,249 test images scored by six metrics that average into relevance and factuality aspects), our best system, 7B-QDoRA, placed fifth out of the participating teams. This rank sits inside a narrow top cluster: the top five teams span roughly 0.03 in overall score, and our BERTScore, the official primary metric, is among the highest in the field, so we read the result as competitive top-cluster captioning rather than a mid-field outcome. On the radiologist-judged explainability subtask, the same system family produced the top result, winning six of nine evaluation criteria, and the radiologist singled out the 7B-QDoRA captions as the most clinically accurate among the top 3 results. The two evaluations agree at the system level (7B-QDoRA is our strongest system on the automated metrics and the one the radiologist preferred), but they diverge at the level of how our team stands relative to the field: mid-cluster on the automated leaderboard, yet field-leading under expert clinical assessment. The two assessments are not directly comparable (the explainability set is a subset of the captioning test images, scored by a clinician rather than by automated metrics), but the contrast is the most interesting observation in this work, and it is only visible because we entered both subtasks.

This paper makes three contributions. First, we report a controlled four-system study that isolates the effect of model scale (0.5B versus 7B) and adapter formulation (QLoRA versus QDoRA) on radiology caption quality. Second, we describe a model-intrinsic, attention-based explainability pipeline whose layer-characterization step separates grounding attention from register and attention-sink behavior, in contrast to the common heuristic of averaging attention over the final decoder layers. Third, we analyze the divergence between automated-metric ranking and radiologist judgment that arises from entering both subtasks. The remainder of the paper reviews related work (Section 2), details the shared methods (Section 3), reports caption-prediction and explainability results (Sections 4 and 5), discusses the cross-task findings and limitations (Section 6), and outlines future work (Section 7).

2. Related Works

Vision-language captioning of medical images: Image captioning has moved from task-specific encoder-decoder models toward general vision-language models that pair a vision encoder with a language model and are adapted to a target domain. Adapting such models to medicine is an active line of work; LLaVA-Med, for example, instruction-tunes a general vision-language model on biomedical figure-caption pairs [7]. We build on LLaVA-OneVision [3], which couples a SigLIP vision encoder [8] with a Qwen2 language model [9] and processes images at native resolution through an anyres tiling scheme. The radiology setting is the one targeted by the ImageCLEFmedical caption task, whose data derive from ROCov2 [4], a large collection of radiology images paired with reference captions and concepts; we fine-tune on that data rather than train a captioner from scratch.

Parameter-efficient fine-tuning: Updating every weight of a multi-billion-parameter model is costly, so a line of methods trains only a small number of added parameters instead. LoRA [10] injects low-rank update matrices into the frozen weights; QLoRA [5] makes this feasible on a single accelerator by quantizing the base model to 4-bit while training the adapter in higher precision; and DoRA [6] decomposes each weight into a magnitude and a direction and applies the low-rank update to the direction, recovering some expressiveness lost to the low-rank constraint. We compare the quantized forms of the latter two, QLoRA and QDoRA, at two model scales.

Attention as explanation: Reading a transformer’s attention weights as an indication of what the

model uses is an established interpretability practice [11], from per-layer per-head visualization [12] to analyses showing that individual layers and heads specialize [13]. Whether attention constitutes an explanation has been debated: one argument is that high-attention positions are not reliably the ones responsible for a prediction [14], and a rebuttal shows that, under controlled comparison, attention does carry explanatory signal [15]. Later work responds to the critique by aggregating attention across layers, as in attention rollout and attention flow [16], or by replacing raw attention with gradient-based relevance propagation through the transformer [17]. A complementary finding is that not all attention mass is semantic: transformers allocate attention to fixed sink positions [18], and vision transformers develop register tokens that hold global information unrelated to any particular image region [19]. Our pipeline is built with these findings in mind: rather than averaging attention over the final decoder layers, the common default, we identify and separate grounding-like attention from register-like and sink-like attention, so that the regions we visualize reflect where the model attends to image content rather than non-semantic attention mass.

3. Methods

The same four systems address both subtasks. This section describes the data (Section 3.1), the model family and its architecture (Section 3.2), the visual token layout that constrains the explainability analysis (Section 3.3), the parameter-efficient fine-tuning (Section 3.4), caption generation (Section 3.5), the attention-based explainability pipeline (Section 3.6), and the compute and reproducibility details (Section 3.7).

3.1. Data

All systems are trained and evaluated on updated version of ROCov2 [4], the dataset adopted by ImageCLEFmedical 2026 for both subtasks [1]. ROCov2 is an updated multimodal collection of radiology figures drawn from the PubMed Central Open Access subset, each paired with a caption and UMLS concept metadata, spanning a broad range of modalities including chest X-ray, CT, MRI, ultrasound, echocardiography, angiography, mammography, retinal fundus, dental panoramic, microscopy, and pathology. We use the partition fixed by the task organizers, which is determined by image-ID prefix rather than by a random split: a captioned development pool of 116,604 images divided into 97,364 training and 19,240 validation images, a separate test set of 15,249 images for the caption-prediction subtask (released without reference captions), and a separate set of 8 images for the explainability subtask. The validation images additionally serve as the sampling pool for the attention diagnostics described in Section 3.6.

3.2. Models and architecture

Our systems are built on the LLaVA-OneVision-Qwen2 family [3], instantiated at the 0.5B and 7B parameter scales. Each model couples a SigLIP-SO400M vision encoder [8] to a Qwen2 language decoder [9] through a two-layer multimodal projector. The vision encoder is identical at both scales: it encodes a 384×384 crop into a 27×27 grid of 729 patch tokens (one token per 14×14 -pixel patch), and this fixed layout is what allows the same heatmap and bounding-box machinery to apply unchanged at either scale. Only the language decoder grows with scale. Table 1 lists the architectural parameters, read directly from the model configurations.

Following Qwen2’s design, the family scales primarily by widening rather than deepening: from 0.5B to 7B the hidden size grows fourfold and the head count doubles, while the decoder depth increases only modestly, from 24 to 28 layers. This has a direct methodological consequence for Section 3.6: a layer that exhibits a particular attention behavior at one scale has no fixed counterpart at another scale by absolute index, so the attention diagnostic is applied to each model independently rather than transferred between them.

Table 1

Architecture of the two model scales. The SigLIP-SO400M vision encoder is shared.

Specification	0.5B	7B
Decoder layers	24	28
Attention heads	14	28
Key/value heads (GQA)	2	4
Hidden size	896	3584
Intermediate (FFN) size	4864	18944
Head dimension	64	128
Vocabulary size	152,000	152,128
Maximum position embeddings	32,768	32,768

3.3. Visual token layout and the base-crop restriction

LLaVA-OneVision uses an anyres scheme to process non-square and high-resolution images without discarding detail through downsampling. The input image is matched to the best-fitting grid from a fixed set of pinpoints (all multiples of the 384-pixel crop size), resized and padded to that grid, and tiled into 384×384 crops, each of which yields 729 patch tokens. A separate base crop, the whole image resized to a single 384×384 , is encoded as well. The processor concatenates the base crop first, followed by the high-resolution tiles row by row, inserting a learned row-separator token after each row, so the total visual sequence length is

$$\text{total_image_tokens} = 729 + (\text{tile_rows} \times \text{tile_cols}) \times 729 + \text{tile_rows},$$

to which roughly 60 text tokens are added for the prompt and the generated caption. Table 2 shows how this grows with image size; large landscape images can exceed ten thousand visual tokens. We cap the tiling at nine high-resolution tiles for caption generation.

Table 2

Visual token counts under anyres for representative image sizes.

Image dimensions	Grid selected	Tiles	Total image tokens
400×400	384×384	1×1	1,459
1024×768	1152×768	3×2	5,106
1920×1080	1920×1152	5×3	11,669
2304×2304	2304×2304	6×6	26,979

For the explainability analysis we adopt a deliberate restriction, which we call the base-crop restriction: heatmaps are computed only from attention over the first 729 base-crop tokens, which map cleanly onto the 27×27 grid, and attention over the high-resolution tiles is discarded. This keeps the spatial mapping exact, at the cost of patch granularity (roughly 14 pixels in the model’s 384×384 input space, or 18 to 22 pixels in the original image), and it means our heatmaps reflect global-view grounding rather than fine tile-level detail. The same token explosion is the source of a memory ceiling during attention extraction: capturing full attention matrices requires the eager attention backend (Section 3.6), whose memory cost scales with the square of the sequence length, so the longest anyres sequences exhaust GPU memory. We note the practical consequence in Section 3.6.

3.4. Parameter-efficient fine-tuning

We adapt each base model with two parameter-efficient methods so the effect of the adapter formulation can be compared directly. QLoRA [5] attaches low-rank adapters to a base model whose weights are quantized to 4-bit NormalFloat (NF4) with double quantization, reducing memory roughly fourfold relative to half precision while keeping the adaptation in higher-precision low-rank factors. QDoRA

combines the same 4-bit quantized base with DoRA [6], which decomposes each weight update into a learnable magnitude per output column and a LoRA-parameterized direction and is intended to recover more of the expressiveness of full fine-tuning. Both methods use bfloat16 compute.

The four systems share one adapter configuration: rank $r = 16$, scaling $\alpha = 32$ (a ratio of 2.0), and dropout 0.05 [10], applied to eight target modules. Six of these are the Qwen2 decoder’s attention and feed-forward projections (the query, key, value, gate, up, and down projections); the remaining two are the LLaVA-OneVision multimodal projector’s two linear layers, which we include because the projector is where vision-to-language alignment is formed and is therefore central to transferring the model into the radiology domain. The SigLIP vision encoder, the language-model head, and the token embeddings are kept frozen. This yields about 11M trainable parameters for the 0.5B systems (roughly 2.1% of the total) and about 40M for the 7B systems (under 1%). All systems are trained with the single instruction prompt “Describe this medical image.” formatted with the Qwen2 chat template.

3.5. Caption generation

Captions are generated by greedy decoding (no sampling), with a cap of 128 new tokens and the same prompt used in training. Decoding the 15,249 test images with each of the four systems produces the four caption-prediction runs evaluated in Section 4. Greedy decoding makes the outputs deterministic, which we adopt for reproducibility and for consistency across the submitted runs.

3.6. Explainability pipeline

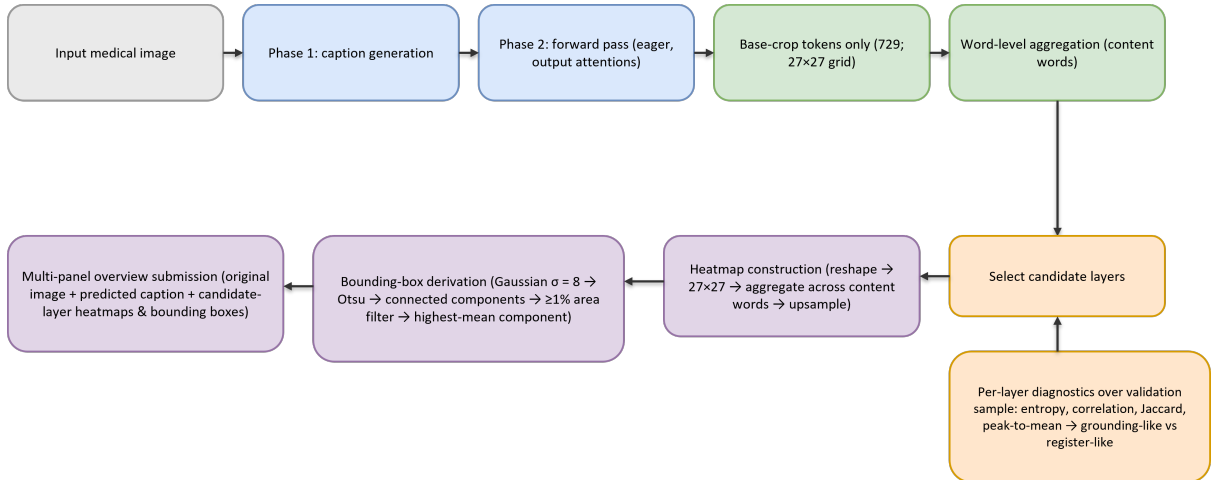


Figure 1: Model-intrinsic explainability pipeline. For each image the captioning model first generates a caption (Phase 1), and a second forward pass with the eager attention backend then exposes per-layer attention (Phase 2). Attention is restricted to the 729 base-crop tokens (the 27×27 SigLIP grid), aggregated over content words, and turned into a summary heatmap and bounding box for each candidate layer. Candidate layers are chosen from a per-layer diagnostic (entropy, correlation, Jaccard, peak-to-mean) that characterizes each decoder layer as grounding-like or register-like. The working-note submission for each image is the multi-panel overview of all candidate layers rather than a single committed layer.

The explanation for a caption is derived from the captioning model’s own attention, in two passes over each image (Figure 1). In the first pass the model generates the caption by greedy decoding. In the second pass the full token sequence (prompt, image tokens, and generated caption) is run through a single forward call with attention and hidden-state outputs enabled. This second pass requires the eager attention backend, because the fused kernels used by default do not return the intermediate attention weights; the eager backend materializes the full attention matrix, which is both slower and far more memory-intensive, and is the reason the longest anyres sequences exceed GPU memory (Section 3.3). From each layer we keep the attention directed from the generated caption tokens onto

the 729 base-crop image tokens. Sub-word attention is averaged within each word, and a stoplist of function words restricts the analysis to content words.

Not every decoder layer’s attention is a useful explanation. Concentrated attention is necessary for grounding but is not sufficient, because a layer can concentrate its attention on a few fixed positions independent of the input, the attention-sink or register behavior documented for autoregressive transformers [18] and for vision transformers [19]. To separate the two, we compute four per-layer diagnostics over a sample of validation images (1,000 sampled with a fixed seed): the mean attention entropy, the mean pairwise correlation between the attention vectors of different content words, the mean overlap (top-10 Jaccard) of their most-attended patches, and the peak-to-mean ratio. Ranking the layers by these measures characterizes each layer as grounding-like (input-dependent, sharply peaked attention that shifts with the words) or register-like (high-correlation, fixed-position attention that is largely independent of the caption). This characterization, rather than a fixed-depth heuristic, is what determines which layers we treat as candidate sources of grounding.

A heatmap is built from a chosen layer by reshaping its 729-dimensional base-crop attention into the 27×27 grid, aggregating across the content words into a single summary map, and upsampling to the image resolution. A bounding box is then derived from the summary heatmap by Gaussian smoothing ($\sigma = 8$ pixels at the upsampled resolution), Otsu thresholding [20], connected-component labeling, discarding components smaller than 1% of the image area, and selecting the component with the highest mean heatmap value, whose bounding rectangle is returned. Selecting by mean rather than by peak or total intensity avoids favoring single-pixel artifacts or merely large regions; by construction the box marks a single region even when the heatmap shows attention in several places.

For the explainability submission we did not commit to a single production layer. Instead, each image’s submission is an investigative multi-panel overview: the original image, the predicted caption, and one panel per candidate layer configuration, each a summary heatmap with its bounding box. Each overview spans candidate layers from both the grounding-like and register-like categories. For instance, the 7B systems, whose overview we use as one of the representative examples (Section 5), the configurations rendered were layers L2, L8, L14, L15, L16, the average of L14 through L16, L22, and the average of L24 through L27; the 0.5B systems used an analogous candidate set identified by the same diagnostic.

3.7. Compute and reproducibility

Hardware: All models were trained on a single node of eight A100 (40 GB) GPUs, using bfloat16 with 4-bit NF4 quantization and gradient checkpointing, with PyTorch DistributedDataParallel across the eight GPUs via torchrun. Caption generation for the 0.5B and 7B systems runs on a single 40 GB A100, since the 4-bit model fits on one GPU, and the explainability captures run on a single A100 or H200 with the eager attention backend.

Training recipe: All four systems use the AdamW optimizer with a 3% warmup ratio and 0.01 weight decay, a per-device batch size of 1 with two gradient-accumulation steps across eight GPUs (an effective batch size of 16), and a learning rate of 2×10^{-4} for the 0.5B systems and 1×10^{-4} for the 7B systems (a lower rate for the larger model follows standard practice). With 97,364 training images, the effective batch corresponds to 6,086 optimizer steps per epoch. Validation loss is evaluated every 500 steps, and training stops early when it has not improved for four consecutive evaluations, after which the best checkpoint by validation loss is restored. All four systems converged and early-stopped within three epochs.

Per-variant compute: Table 3 reports training time, per-step time, peak training memory, the epoch at convergence, the final training loss, and per-image inference latency and memory. Two patterns matter for reproduction. First, QDoRA is substantially more expensive than QLoRA at inference, roughly twice as slow at 0.5B and about three and a half times as slow at 7B, because DoRA recomputes a per-layer magnitude vector; this overhead is the dominant cost when scaling QDoRA to large test sets. Second, training memory is governed by activations and optimizer state rather than by the adapter, so QLoRA and QDoRA at a given scale have nearly identical training footprints. The absolute inference

latencies are post-training probes over 500 validation images and are not directly comparable across scales (the 7B QLoRA probe ran on a single H200); the robust signal is the QDoRA-over-QLoRA overhead within each scale.

Table 3

Per-variant training and inference compute.

Variant	Train time (h)	Time/step (s)	Peak train mem (GB)	Convergence (epoch)	Final train loss	Inference (s/img)	Inference mem (GB)
0.5B QLoRA	10.15	2.61	18.31	2.30	2.407	2.49	8.56
0.5B QDoRA	15.51	3.99	18.38	2.30	2.403	4.73	8.56
7B QLoRA	10.86	4.89	29.26	1.32	2.136	1.63	23.70
7B QDoRA	28.08	7.23	29.61	2.30	2.017	5.63	23.72

Reproducibility. The environment pins transformers 5.3.0, peft 0.18.1, bitsandbytes 0.49.2, scipy 1.17 or later, and PyTorch 2.11.0 (CUDA 12.8). The validation sample for the attention diagnostics is drawn with a fixed seed of 42, using the same indices across model scales, and caption decoding is greedy and therefore deterministic. The explainability captures use the eager attention backend; all other inference uses the default fused backend.

4. Results: Caption Prediction

The caption-prediction subtask is scored by six automated metrics that the organizers group into a relevance aspect and a factuality aspect and then average into a single overall score [1]. Relevance is the mean of four metrics: BERTScore, the official primary metric, together with ROUGE-1, an image-caption similarity score, and BLEURT. Factuality is the mean of two metrics: a UMLS concept F1 computed with MedCAT, and AlignScore. The overall score is the mean of the two aspects. We submitted one run per system.

Table 4

Caption-prediction scores for our four systems on the 15,249-image test set. Best value in each column in bold.

System	Overall	Relevance	Factuality	BERTScore	ROUGE	Similarity	BLEURT	UMLS F1	AlignScore
0.5B-QLoRA	0.3325	0.5247	0.1403	0.5997	0.2597	0.9146	0.3248	0.1593	0.1214
0.5B-QDoRA	0.3317	0.5230	0.1403	0.5978	0.2580	0.9110	0.3254	0.1597	0.1209
7B-QLoRA	0.3343	0.5257	0.1429	0.6022	0.2689	0.9077	0.3239	0.1594	0.1264
7B-QDoRA	0.3458	0.5377	0.1539	0.6075	0.2725	0.9408	0.3299	0.1751	0.1327

7B-QDoRA is our strongest system, posting the best value on every metric in Table 4. Its margin over the next-best system is widest on image-caption similarity (0.9408, about 0.026 ahead) and on UMLS F1

Table 5

Official caption-prediction leaderboard, by overall score. Our entry (rank 5) is shown in **bold**; other teams are anonymized. Per-team figures are from the task overview [1].

Rank	Overall	Relevance	Factuality	BERTScore	ROUGE	Similarity	BLEURT	UMLS F1	AlignScore
1	0.3755	0.5786	0.1724	0.6250	0.3234	1.0104	0.3555	0.2160	0.1287
2	0.3603	0.5527	0.1679	0.6017	0.2725	0.9902	0.3463	0.1822	0.1537
3	0.3571	0.5365	0.1777	0.6042	0.2809	0.9381	0.3229	0.1943	0.1610
4	0.3516	0.5288	0.1743	0.6130	0.2728	0.9113	0.3183	0.1812	0.1674
5	0.3458	0.5377	0.1539	0.6075	0.2725	0.9408	0.3299	0.1751	0.1327
6	0.3343	0.5201	0.1484	0.6013	0.2226	0.9273	0.3293	0.1563	0.1406
7	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379
8	0.3065	0.4705	0.1426	0.5465	0.2224	0.7787	0.3342	0.1443	0.1408
9	0.2990	0.4798	0.1181	0.5670	0.2230	0.8147	0.3146	0.1048	0.1315
10	0.2814	0.4526	0.1102	0.5690	0.2126	0.7286	0.3004	0.1427	0.0778
11	0.2685	0.4346	0.1025	0.5398	0.2059	0.6788	0.3137	0.1219	0.0831
12	0.2673	0.4275	0.1072	0.5742	0.1818	0.6589	0.2951	0.1155	0.0989
13	0.2527	0.4373	0.0680	0.5720	0.2207	0.6514	0.3051	0.0872	0.0488

(0.1751, about 0.015 ahead), with a smaller but consistent edge on the remaining metrics.

Against the field, 7B-QDoRA placed fifth out of the participating teams (Table 5), but the bare rank understates the outcome. The leading systems are tightly bunched, with the top five separated by roughly 0.03 in overall score. Our standing rests on the metrics where we are strongest. Our BERTScore (0.6075) ranks third in the field, and our image-caption similarity (0.9408) is likewise third, placing us among the strongest systems on both. Our weaker axes are factuality (0.1539), where we trail the other top-cluster teams, and AlignScore (0.1327). Both point to room for improvement in claim-level faithfulness rather than surface overlap. We therefore characterize the result as competitive top-cluster captioning rather than a mid-field placement. One figure in the released results should be read with caution. The top-ranked system’s image-caption similarity (1.0104) exceeds the range the metric is expected to occupy, which suggests a scoring artifact, so the margin at the very top of the leaderboard should not be over-interpreted.

5. Results: Explainability

The explainability subtask placed no technical restriction on the form of the explanation and was assessed by a radiologist on a five-point scale across nine criteria, with an additional overall score [1]. For each evaluation image we submitted the investigative multi-panel overview described in Section 3.6. Table 6 reports the radiologist’s scores for the top 3 results.

Table 6

Radiologist Likert scores (1 to 5) on the explainability subtask, ranked by overall score. Our entry is rank 1; the other results are anonymized. Best value in each column in bold.

Rank	Readability	Accuracy	Level of Detail	Caption focus	Consistency	Comprehensiveness	Visual focus	Methodology	Clinician’s favorite	Overall
1	4.6	3.6	3.4	4.3	3.1	2.9	3.4	3.0	5.0	3.69
2	2.63	2.13	2.75	2.63	2.63	3.13	2.75	4.0	3.0	2.85
3	4.5	3.1	3.6	4.1	1.0	2.1	1.0	3.0	3.0	2.83

Our submission was the top result, with an overall score of 3.69; the next two results were 2.85 and 2.83, a margin of about 0.85 points. We obtained the highest score on six of the nine criteria: readability, accuracy, caption focus, visualization consistency, visualization focus, and the clinician’s favorite, the last a maximum of 5.0. We trailed on three criteria: level of detail (narrowly, behind one other result), comprehensiveness, and methodology; the methodology score is taken up in Section 6.

The radiologist judged our captions stylistically polished and, in particular, identified the 7B-QDoRA variant as producing the most clinically accurate findings among the top 3 results. This is consistent with 7B-QDoRA being our strongest system on the automated metrics (Section 4): automated quality and clinical judgment agree on which of our systems is best. The radiologist scored each team’s submission per image rather than per individual system. Figures 2 and 3 show, as representative examples, our submitted explainability overviews on test image 10741 (coronary angiography), the 0.5B systems in Figure 2 and the 7B systems in Figure 3. This is an easy angiography case on which our submission was the top result: a per-image mean of 3.86 across the seven per-image criteria, with the next two results at 2.71 and 2.14. Each overview comprises the original image, the predicted caption, and the candidate-layer summary heatmaps with their bounding boxes. Table 7 lists the ground-truth reference caption and the four systems’ predicted captions for this image.

Across the scored images our submission was the radiologist’s top-scored result on seven of the eight images that were evaluated. Our strongest image was a difficult CT case (test_4294, per-image mean 4.00), and our single loss was an easy CT case (test_6753, 3.14), which received our lowest level-of-detail rating. The pattern that our larger margins tended to fall on the cases rated more difficult suggests the approach is at least robust to image difficulty on this small evaluation set.

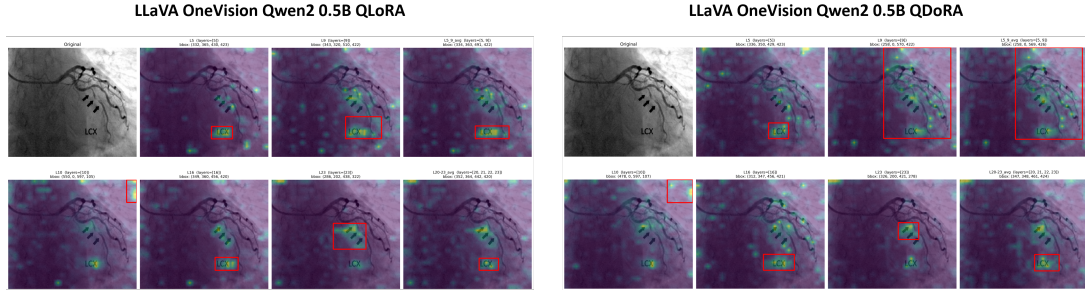


Figure 2: Original image attribution: CC BY [Abdikarimov et al. (2024)]. Representative submitted explainability overviews for the 0.5B systems (QLoRA and QDoRA) on test image 10741 (coronary angiography). Each overview shows the original image, the predicted caption, and one summary heatmap with a derived bounding box per candidate layer.

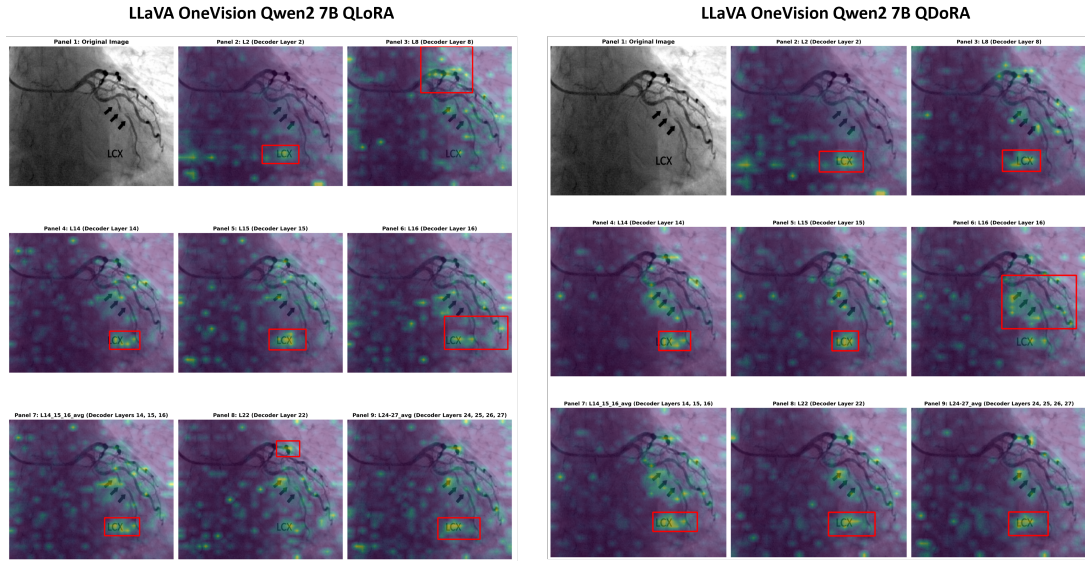


Figure 3: Original image attribution: CC BY [Abdikarimov et al. (2024)]. Representative submitted explainability overviews for the 7B systems (QLoRA and QDoRA) on test image 10741, in the same format as Figure 2.

6. Discussion

6.1. Automated metrics versus radiologist judgment

Entering both subtasks places our systems in two positions relative to the field: mid-cluster on the automated caption benchmark, where our best system ranked fifth out of the participating teams inside a tight top cluster (Section 4), and field-leading on the radiologist-judged explainability subtask, where the same system family produced the top result, the top 3 results being 3.69, 2.85, and 2.83 (Section 5). The two are not directly comparable, since the explainability set is a small subset of the captioning test images scored by a clinician’s holistic judgment rather than by automated overlap with reference captions. They are, however, mutually reinforcing on the question they share, namely which of our systems is best: the automated metrics and the radiologist independently single out 7B-QDoRA, the larger model with the DoRA adapter, as our strongest system. An automated benchmark over the full test set and an expert reading of a small subset converging on the same system gives us more confidence in 7B-QDoRA than either evaluation would alone.

Table 7

Ground-truth reference caption and the four systems’ predicted captions for test image 10741.

Source	Caption
Ground truth	Coronarography. LCX: distal left circumflex coronary artery stenosis 95%.
0.5B-QLoRA	Coronary angiography showing the left circumflex artery (LCX) with a 90% stenosis (arrows) in the mid-segment.
0.5B-QDoRA	Coronary angiography showing the left circumflex artery (LCX) with multiple collateral vessels (arrows) arising from the left circumflex artery.
7B-QLoRA	Coronary angiography of the left coronary system showing a large number of collateral vessels (arrows) arising from the left circumflex coronary artery.
7B-QDoRA	Left coronary angiography in shows a tortuous left circumflex artery (LCX) with multiple small aneurysms (black arrows).

6.2. Limitations

Our heatmaps are computed from attention over the 729 base-crop tokens only (Section 3.3), because the base crop maps exactly onto the 27×27 SigLIP grid and keeps the attention-to-region correspondence unambiguous, whereas the tiled anyres tokens carry no comparably clean global mapping. The restriction proved effective: the base-crop explanations produced the top result overall and won the visualization-focus criterion (Section 5), so base-crop attention was sufficient to localize clinically relevant regions for an expert reader. It nonetheless fixes the heatmaps at a patch granularity of roughly 14 pixels and sets aside whatever caption-relevant attention the high-resolution tiles carry.

Three further limitations bound the explanations. The heatmaps are read directly from decoder attention, which we treat as an indication of where the model looks but have not validated against gradient-based or attention-rollout faithfulness measures. The bounding box per layer is a single thresholded region, suited to a focal finding but not a multi-focal or bilateral one. And the radiologist evaluation covers only a small set of images, so the per-criterion patterns we report are indicative rather than statistically established.

7. Future Work

The follow-up to this submission replaces the investigative overview with a single committed visualization: for each image, one summary heatmap and bounding box drawn from the selected grounding layer, together with per-word heatmaps and bounding boxes for the individual findings, which target the consistency, focus, and comprehensiveness criteria more directly than the overview did. Beyond the output format, we plan to validate the attention-based explanations against gradient-based and attention-rollout faithfulness measures, to extend the heatmaps from the base crop to the high-resolution anyres tiles, and to characterize the grounding, register, and attention-sink behavior of the decoder layers across model scales, including the 72B model that fell outside the submission window.

8. Conclusion

We addressed both ImageCLEFmedical 2026 subtasks with a single family of vision–language models, LLaVA-OneVision-Qwen2 at the 0.5B and 7B scales, each adapted to radiology with QLoRA and QDoRA. Our strongest system, 7B-QDoRA, placed fifth out of the participating teams on the automated caption benchmark inside a narrow top cluster, and a model-intrinsic attention pipeline turned the same model into its own explainer, producing the top result on the radiologist-judged explainability subtask and winning six of nine criteria. Although our standing relative to the field differs between the two leaderboards, the evaluations reinforce each other where they overlap: both the automated metrics and the radiologist identify 7B-QDoRA as our strongest system, which gives us confidence in that choice going forward.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and other generative AI tools solely for grammar correction, spell-checking, and minor language polishing to improve the readability and clarity of the manuscript. After using these tools, the authors carefully reviewed and edited the resulting text as needed and take full responsibility for the content of this publication.

Acknowledgements

This work used Delta GPU at the National Center for Supercomputing Applications (NCSA) through allocation CIS251271 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [21], which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

- [1] 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. To appear.
- [2] 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.
- [3] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, C. Li, LLaVA-OneVision: Easy visual task transfer, arXiv preprint arXiv:2408.03326 (2024).
- [4] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. 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.
- [5] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.
- [6] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, DoRA: Weight-decomposed low-rank adaptation, in: ICML, 2024, pp. 32100 – 32121.
- [7] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day, in: NeurIPS, 2023.
- [8] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11975–11986.
- [9] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, et al., Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

- [10] 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: International Conference on Learning Representations (ICLR), 2022. ArXiv:2106.09685.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), 2017. arXiv:1706.03762.
- [12] J. Vig, A multiscale visualization of attention in the transformer model, in: ACL, 2019, pp. 37 – 42. doi:10.18653/v1/p19-3007.
- [13] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of BERT's attention, in: BlackboxNLP@ACL, 2019, pp. 276 – 286. doi:10.18653/v1/w19-4828.
- [14] S. Jain, B. C. Wallace, Attention is not explanation, in: NAACL-HLT, 2019, pp. 3543 – 3556. doi:10.18653/v1/n19-1357.
- [15] S. Wiegrefe, Y. Pinter, Attention is not not explanation, in: EMNLP/IJCNLP, 2019, pp. 11 – 20. doi:10.18653/v1/d19-1002.
- [16] S. Abnar, W. H. Zuidema, Quantifying attention flow in transformers, in: ACL, 2020, pp. 4190 – 4197. doi:10.18653/v1/2020.acl-main.385.
- [17] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791. doi:10.1109/CVPR46437.2021.00084.
- [18] G. Xiao, Y. Tian, B. Chen, S. Han, M. Lewis, Efficient streaming language models with attention sinks, in: ICLR, 2024.
- [19] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision transformers need registers, in: International Conference on Learning Representations (ICLR), 2024. ArXiv:2309.16588.
- [20] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1979) 62–66.
- [21] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, J. Towns, ACCESS: Advancing innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services and support, in: Practice and Experience in Advanced Research Computing (PEARC '23), ACM, New York, NY, USA, 2023. doi:10.1145/3569951.3597559.