

# RAD-ACE: MULTIMODAL LARGE LANGUAGE MODELS AS RADIOLOGY ASSISTANTS

**Can A. Ateş, Abdullah E. Ergün, Emre Çoban**

Department of Computer Science

Hacettepe University

Ankara, 06800, Turkey

{canaliateş, ergunenes7, emrecobann02}@gmail.com

Github: <https://github.com/canatess/RAD-ACE>

## 1 INTRODUCTION

In medical domain, radiology plays an important role by providing reports for clinical decisions. These radiology reports which summarize medical findings from radiology images, bridge the gap between raw visual data and actionable clinical insights. However, the generation of these reports is a very labor-intensive process that relies heavily on the radiologist's expertise and workload that makes the process both time-consuming and susceptible to errors. In addition to these vulnerabilities, especially after COVID-19 pandemics, demand for radiology reports is increasing exponentially [1]. In the field of radiology, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans are the most demanding types among all because of their ability to reveal fine anatomical structures and pathological patterns. Interpreting these images requires attention, domain expertise, and structured reasoning, which significantly increases the cognitive load on radiologists. The complexity and volume of these scans often lead to longer turnaround times and greater variability in reporting quality, especially when handled under pressure. Because of these challenges, there is a growing need for smart systems that can help create radiology reports automatically or with minimal help from humans. These systems should be able to produce accurate, clear, and reliable reports that match what doctors expect in real medical settings. Solving this problem is important to make the reporting process faster, reduce delays in diagnosis, and keep the quality of patient care high in all types of hospitals and clinics.

Recent advancements in artificial intelligence have opened new possibilities for addressing the challenges of radiology report generation. In particular, multimodal large language models (MLLMs)—capable of understanding both images and text—offer a powerful framework for bridging visual perception and clinical reasoning. These models can be trained to analyze medical images, follow diagnostic instructions, and produce structured, interpretable reports that align with real-world clinical needs. By mimicking the step-by-step reasoning process of expert radiologists, MLLMs not only improve the consistency and speed of reporting but also enhance transparency and trust in automated systems. As a result, they present a promising path toward scalable, intelligent assistants that can ease the burden on radiologists while supporting high-quality, timely patient care.

This study introduces RAD-ACE, which is a MLLM framework that adapts state-of-the-art open-source MLLMs for use as radiology assistants. This framework is designed to follow expert-like clinical reasoning to generate detailed and interpretable reports directly from medical images. By leveraging recent advancements in vision-language modeling, the framework aims to automate the radiology reporting process in a way that is transparent, scalable, and aligned with real-world diagnostic practices. Based on these problems and motivation, the study contributes to:

1. **RAD-ACE-CoT** which is a novel dataset, contains radiology images paired with chain-of-thought diagnosis to determine clinical findings.
2. **RAD-ACENTS** which are radiology-specialized agents with serviceable capability of report generation from radiology images.

## 2 TECHNICAL

In this section, technical aspects of the study are explained. Section 2.1 explains the preparation of **RAD-ACE-CoT** data. Section 2.2 examines the technical details of the models are used. Section 2.3 discusses the overall methodology employed for model training and evaluation.

### 2.1 RAD-ACE-CoT DATA PREPARATION

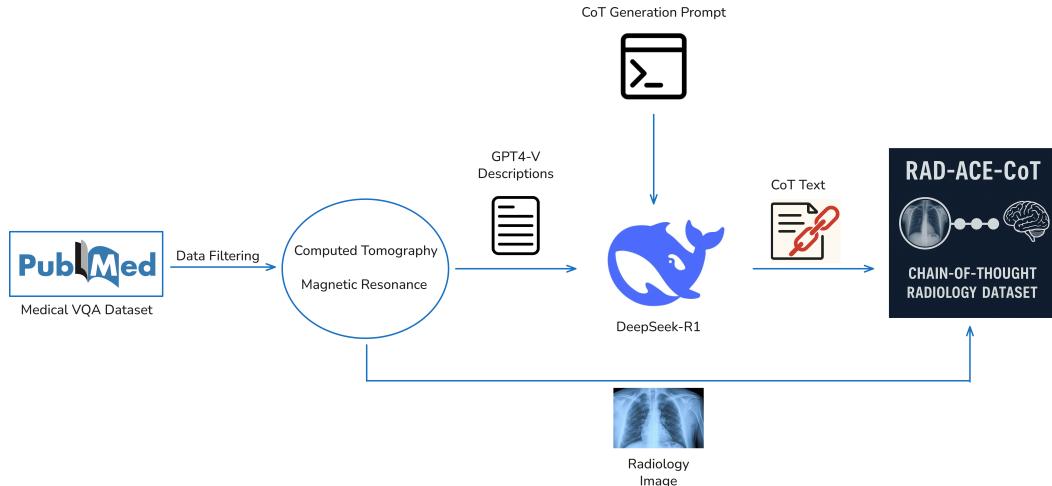


Figure 1: RAD-ACE-CoT Generation

The RAD-ACE-CoT, a novel multimodal chain-of-thought radiology dataset, is the first contribution of the study. As illustrated in Figure 1, the data generation pipeline comprises two key steps.

As a first step, the **PubMedVision** dataset [2]—our primary source, containing approximately 1.6 million high-quality, large-scale medical VQA samples—was filtered to enhance its usability for this study. In this filtering, due to computational and economical reasons, a subsample that contains 12K Computed Tomography (CT) and Magnetic Resonance (MR) images with related GPT4-V [3] generated descriptions are selected.

In second step, a study-unique prompt A.1 which aims to generate Chain-of-Thought process from the visual descriptions is developed. This developed prompt is used in DeepSeek-R1 [4] to generate CoT-texts from the GPT4-V generated descriptions in selected subsample. These CoT-texts and their related images which in the subsample is matched to build up final RAD-ACE-CoT dataset. Consequently, the our dataset consists 12K high-quality image-CoT pairs to generate a medical report.

### 2.2 MODELS

Given that the computer vision problem addressed in this study lies at the intersection of visual and textual modalities, multimodal large language models are employed due to natural capable of processing both image and text data. **3B - 7B** versions of **Qwen2.5-VL** [5,6], and **11B** version of **LLAMA-3.2-Vision** [7,8,9] models were selected for this purpose. One of the main limitations are arised here, there are very limited open-source model pool in the field of VLM so number of available models are restricted. The primary motivation behind these selections is to enable a comparative analysis of model capacities and to observe performance differences across distinct architectural configurations.

#### 2.2.1 QWEN-2.5VL

**Qwen2.5-VL** is a general-purpose vision-language model (VLM) architecture developed by Alibaba DAMO Academy [10] for large-scale multimodal understanding. Beside to its predecessors,

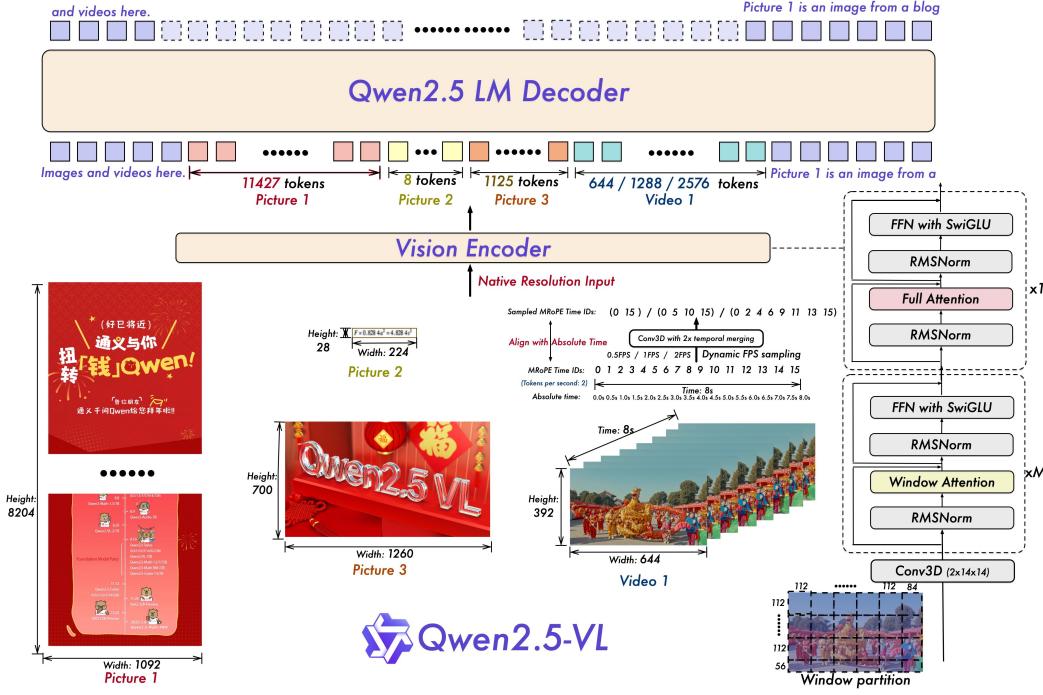


Figure 2: QWEN2.5-VL Overall Architecture

this model family shows significant advancements for cross-modal reasoning by integrating a custom high-capacity Vision Transformer encoder with a Qwen2.5-based decoder-only language model [11]. In this study, 3B and 7B versions of the QWEN2.5-VL family is used because they are most demand versions in terms of model complexity and computational efficiency balance which making them highly applicable for downstream tasks such as image captioning.

**Input** As demonstrated in Figure 2, model family accepts multimodal inputs consisting of visual context and textual prompts. Qwen2.5VL’s vision encoder supports different image resolutions (e.g. 1092x8204, 1260x700, 512x384, 360x720, 224x288) without any resizing operations, and videos captured at their original dimension as a sequence of frames (e.g. 8x644x392x3). The native resolution maintaining property provides preserving spatial details absolutely which makes model more precise at downstream tasks.

**Vision Encoder** As demonstrated in Figure 2, Vision Encoder is the component that placed in center to bridge gap between visual and textual inputs. Researchers do not specify any unique name for it but it could be think like ***custom-designed dynamic-resolution Vision Transformer (ViT)***. With the help of this encoder, model family could be processing varying sizes without resizing. Compared to predecessors (e.g. Qwen1.0-VL, QwenVL-Plus [12]) this version utilizes a window-based self attention mechanism which reduces the standard self-attention’s computational complexity while maintaining rich local interactions.

**PATCH PARTITIONING** In terms of images, each image is dividing into non-overlapping patches which are 14x14 pixel/patch commonly. As an example, a 512x384 image yields 36 (width) x 27 (height) in total 972 patches. Each patch is flattened and linearly projected to a patch embedding form.

On the other hand, videos have one more dimension which is temporal. The vision encoder is extended with adjacent frame grouping technique (e.g. three frame per patch) to accept sequences of frames. The technique mainly based on temporal-aware tokenization, where 3D positional encodings are applied to preserve spatiotemporal order. This protection enables the model to reason across time when processing dynamic visual content.

**ATTENTION** As demonstrated with an extraction in Figure 2, vision encoder uses two different attention modules which are Window Attention [13] ( $M$  times) and Full Attention [14] (1 time).

In window-attention, patch embeddings are grouped into fixed-size (e.g. 112x112) non-overlapping windows where each window contains a fixed grid of patches (e.g. 8x8 grid of tokens). Self-attention is computed independently within each window, meaning that tokens that belong to one window do not attend other windows' tokens. With this strategy, computational cost is decreased from quadratic ( $O(N^2)$ ) to linear ( $O((n/w)^2 \times w^2)$ , where  $n$  is the total number of tokens and  $w$  is the number of tokens per window). Model becomes suitable for processing high resolution inputs while maintaining spatial fidelity in computationally efficient manner with the help of this attention localization. This property makes model be able to process compromising fine-grained details inputs like radiological scans.

In full (global) attention, Qwen2.5-VL introduces self-attention in selected transformer layers across all visual tokens. The interaction between these visual tokens enables capturing global dependencies. Although this is a more computationally expensive operation, applying it sparingly to only selected transformer layers, model benefits from global consciousness without incurring excessive cost. In radiology, these global dependencies are essential to capture contextual cues that span across distant spatial regions such as the bilateral comparison of organs or the identification of multi-region anomalies.

Qwen2.5-VL model family maintains both high spatial fidelity and long-range coherence with this hybrid attention strategy which combines power of windowed and full attention. Empirical evidence from the Qwen2.5-VL technical report [5] indicates that this design significantly boosts performance on downstream tasks involving dense visual reasoning and localization.

**Multimodal Rotary Position Embedding** As a complementary operation, M-RoPE (Multimodal Rotary Position Embedding) [15,16] is applied after patching and embedding to ensure both spatial structure and motion continuity are preserved. Each visual token is enriched with 2D rotary positional embeddings to encode its row and column position in the image grid. On the other hand during processing video frame sequences, M-RoPE extends to 3D positional encoding, that captures the spatial and temporal position of each token with IDs.

**Cross-Modality Adapter** Qwen2.5-VL uses a MLP-based Cross Modality Adapter before feeding visual tokens into the language model. In this adapter; output of the vision encoder, typically a long visual token sequence, is projected to embedding space of language via trainable cross-modal adapter. By utilizing cross-attention, cross-attention compresses and aligns the visual representations into a fixed-length set of multimodal tokens (default length: 256). These compressed tokens form the visual prefix and are prepended to the token stream received by the language decoder.

The fusion design used in this adapter ensures that the conditions of model use for language generation is not restricted with only on the text input but also on detailed visual information, without overwhelming its attention mechanism. The cross-attention adapter and compression strategy enable Qwen2.5-VL to scale to large image resolutions without degrading performance.

**Qwen2.5 LM Decoder** As demonstrated in Figure 2, text generation backbone (decoder) of the Qwen2.5-VL depends on Qwen2.5 Language Model [11] which is a high-performance decoder-only language model. This model employs SwiGLU activation and RMSNorm layer normalization as done in vision encoder for faster training. Also, it uses RoPE (Rotary positional encodings) [17] for text tokens to ensure consistent representation across modalities. The model can generate various formatted NLP responses such as structured outputs like JSON, HTML etc.

**Summary** Qwen2.5-VL integrates a custom dynamic-resolution ViT encoder with a Qwen2.5 decoder-only language model. It accepts native-resolution images and videos, dividing them into 14x14 patches, and applies 2D or 3D positional encodings using M-RoPE to preserve spatial and temporal structure. A hybrid attention mechanism—windowed attention for local efficiency and sparse global attention for contextual understanding—is employed in the vision encoder. Both encoder and decoder use SwiGLU [18] and RMSNorm to enhance training dynamics. A cross-modal adapter compresses visual tokens into a fixed-length prefix, which is prepended to the decoder's input for multimodal generation.

### 2.2.2 LLaMA 3.2: VISION

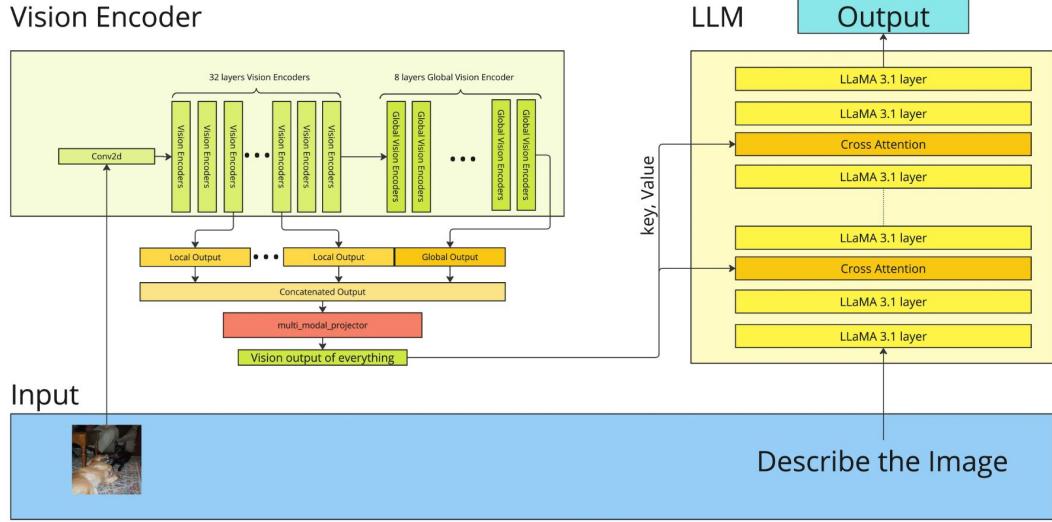


Figure 3: LLaMA3.2-Vision Overall Architecture

As a Meta AI’s flagship vision-language model (VLM) under the LLaMA 3 [7] family, LLaMA 3.2- Vision (MLLaMA) is designed for high-performance multimodal reasoning. Compared to its text-only counterparts, LLaMA 3.2-Vision incorporates a dedicated visual processing pathway and a cross-modal fusion mechanism into the LLaMA 3.1 language model [19]. It supports a wide range of tasks, including image captioning, visual question answering (VQA), document parsing, and chart understanding. In this study 11B version is used which is pretrained on a combination of language-only and image-text datasets.

**Overall Design** In general, LLaMA 3.2-Vision extends the LLaMA 3.1 text-only language model with a multimodal interface. This interface is built upon a two-stage visual encoder, cross-attention-based visual adapter, and LLaMA3.2 decoder [7]. The model can extract visual dense representations, transforms visual embeddings into a compatible form with the language decoder, and can optimize text-centric autoregressive output generation.

**Two-Stage Vision Encoder** The vision encoder in LLaMA 3.2-Vision takes high-quality images and turns them into smaller chunks of useful information (called embeddings). These chunks can then be easily combined with the regular text-processing part of the LLaMA model so it can understand both pictures and words together. This two-stage design with preserved intermediate features enables multi-scale visual understanding, helping the model capture both fine details and abstract patterns. It ensures information retention by saving key features that could be lost in deeper layers. Through gated attention, the model controls how different feature levels are combined. The high-resolution input further enhances the encoder’s ability to capture detailed visual information.

**IMAGE PROCESSING** The LLaMA 3.2-Vision model resizes each image to 448×448 pixels and segments it into non-overlapping patches, each measuring 14×14 pixels. This results in a grid of 32×32 patches, totaling 1,024 visual tokens. Each token is embedded into a 1,280-dimensional feature vector. Compared to common setups like CLIP-ViT-L/14-336 [20], which yields a 24×24 patch grid from a 336×336 input, MLLaMA operates at a higher spatial resolution. This denser patching strategy enhances fine-grained perception and supports more detailed visual reasoning tasks.

**32-LAYERS VISION ENCODERS** All patches are globally processed in these layers which is a similar strategy in standard ViT [21]. One important idea in the design is how it keeps useful image information as it processes the input. According to its settings, the model saves the outputs from layers 3, 7, 15, 23, and 30 of the image encoder. These saved outputs capture features at different levels—from basic shapes to more abstract patterns. By holding onto these layers, the model builds

a mix of low-level and high-level details, helping it better understand the image and perform more advanced tasks.

**8-LAYERS GLOBAL VISION ENCODERS** The 8-layer global encoder introduces two main innovations that improve how visual information is handled and integrated. First, it uses Gated Attention Mechanisms [22]—trainable components that act like filters—allowing the model to control which parts of the image features to focus on or ignore. This selective attention helps it emphasize the most relevant visual cues. Second, the global encoder doesn’t work in isolation; it also connects to the outputs of earlier layers that were saved during image processing. This design enables the model to bring together both detailed low-level features (like edges or textures) and high-level semantic patterns (like objects or scenes). As a result, the model builds a rich, layered understanding of the image, supporting better performance in tasks that require both fine detail and overall visual context.

**FEATURE CONCATENATION** LLaMA3.2-Vision combines the output of the global encoder with five intermediate features from earlier layers, resulting in a single vector of 7680 dimensions (6 feature sets  $\times$  1280 dimensions each). This merged feature captures a wide range of visual information—from low-level details like edges and textures to high-level concepts such as objects and scenes. By providing the language model with this rich, multi-scale representation, the system gains a more complete understanding of the image, which helps improve the quality of generated text. Figure 3 illustrates this two-stage feature extraction process.

**Multimodal Projection Layer** The MMPL acts as a translator between vision and language. It takes a rich 7680-dimensional visual vector—formed by combining the global encoder output with intermediate features—and maps it into a 4096-dimensional space that the language model can understand. This step not only reduces the size but more importantly aligns visual features with the language model’s semantic space, enabling smooth and meaningful interaction between visual input and text generation.

**CROSS-ATTENTION** The module is responsible for integrating visual information into the language generation process. Implemented via MllamaCrossAttentionDecoderLayer, these layers transform the visual encoder’s output into cross\_attention\_states, which are then used as keys and values in the MllamaTextCrossAttention module. During this interaction, text-derived queries attend to visual features, with gating mechanisms modulating the flow of information and separate normalization layers ensuring training stability. This setup enables the model to selectively incorporate relevant visual context throughout the decoding process.

**MULTI-POINT INTEGRATION** In this strategy, cross-attention layers are used every 5 blocks in the language model decoder. This setup enables progressive refinement, where early, middle, and late layers incorporate visual context at increasing semantic depth. By reusing the same projected visual features across layers, the model improves efficiency without redundancy. Gated mechanisms and attention masks ensure flexible, controlled integration, allowing the model to modulate visual influence based on context. This design preserves both rich visual semantics and generation fluency, forming the backbone of MLLaMA’s cross-modal capabilities.

**Language Model** The language decoder of LLaMA3.2-Vision builds upon the 40-layer LLaMA 3.1 architecture with a 4096-dimensional hidden size. To enable multimodal capabilities, it incorporates cross-attention layers at regular intervals—every 5 layers (i.e., layers 3, 8, 13, 18, 23, 28, 33, and 38)—which allow integration of visual features into the language stream. These cross-attention layers, trained via adapter modules on text-image pairs, connect the frozen LLaMA 3.1 weights with a vision encoder. While the self-attention layers remain unchanged from LLaMA 3.1, only the cross-attention and vision encoder components are updated during training to preserve the language model’s pre-trained capabilities.

**Summary** LLaMA 3.2-Vision (MLLaMA) extends LLaMA 3.1 with a two-stage vision encoder and cross-modal fusion for tasks like VQA and image captioning. Its high-resolution patching and gated attention enable rich, multi-scale visual understanding. By integrating visual features through cross-attention layers, it achieves efficient multimodal reasoning while preserving the strengths of the original language model.

### 2.3 METHODOLOGY

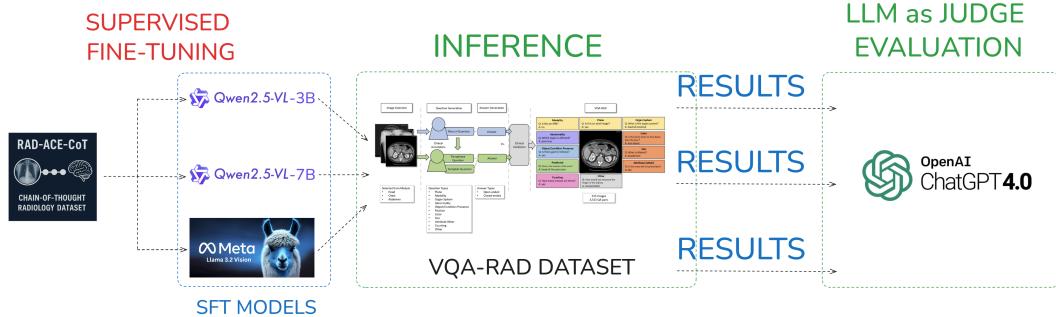


Figure 4: Overall Methodology

Overall methodology for the study is demonstrated in Figure 4. The methodology mainly based on three key steps as follows:

**Supervised Fine-Tuning (SFT)** Qwen2.5-VL-3B, Qwen2.5-VL-7B, and LLaMA3.2-Vision-11B models are supervised fine-tuned using the RAD-ACE-CoT dataset mentioned in in Section 2.1. This process enables the models to adapt to the specific domain of radiology by learning from high-quality image–text pairs annotated with diagnostic reasoning. Through SFT, the models are guided to align their visual and textual representations with clinically relevant patterns, improving their ability to generate accurate, coherent, and chain-of-thought style medical reports. It also enhances their generalization in downstream vision-language tasks by reinforcing task-specific reasoning and structured output generation.

**Inference** In this stage, 100 hard examples are selected from the VQA-RAD dataset [23] by manually ensuring diversity across different anatomical regions. The use of these challenging and diverse samples during inference enables a robust evaluation of the supervised fine-tuned models’ generalization capabilities. Specifically, it allows assessment of the models’ performance on previously unseen, complex radiology cases, providing insights into their reasoning accuracy, robustness, and adaptability in real-world clinical scenarios.

**LLM as Judge Evaluation** The use of RadRScore [24] and LLM as Judge [25] methods was planned for evaluation, but since there is no open-source dataset with the ground truth data required for RadRScore, only the LLM as Judge method was used in this study. For judgment, the ChatGPT4-o [26] model from OpenAI, one of the models with the highest capacity currently available, was used. This model evaluated the inferred reports based on the parameters described in Section 2.1 (“Clinical Relevance, Factuality/Hallucination, Reasoning Coherence, Completeness, Final Answer Quality”) and assigned a score to each model.

In summary, the proposed methodology follows a structured pipeline comprising supervised fine-tuning, inference on challenging radiology samples, and evaluation via an advanced LLM-based scoring framework. This design ensures domain adaptation, robust generalization testing, and reliable performance assessment of multimodal models in the context of radiology report generation.

## 3 TRAINING SETUP & EXPERIMENTS

In this section, we describe the fine-tuning setup used to adapt large-scale multimodal models to the RAD-ACE-CoT dataset. All training was conducted using the *Unsloth* framework [27], which provides memory-efficient infrastructure for fine-tuning large vision-language models (VLMs). Our training protocol relies on **supervised fine-tuning (SFT)** with structured image–CoT pairs described in Section 2.1.

### 3.1 MODEL SELECTION AND PRECISION SETUP

We selected three open-source multimodal large language models based on their balance between scale, accessibility, and performance in vision-language reasoning tasks:

- **Qwen2.5-VL-3B** – 16-bit, decoder-only VLM
- **Qwen2.5-VL-7B** – 16-bit, decoder-only VLM
- **LLaMA-3.2 Vision-11B** – 16-bit, large decoder-only VLM

All models were initialized with pretrained weights and then fine-tuned using LoRA [28] adapters (Rank = 16,  $\alpha$  = 16) applied to both vision and language layers. Full LoRA injection was enabled for attention and MLP modules, while no vision layers were frozen during training to allow joint adaptation across modalities. All models were trained in **16-bit precision**, with bf16 enabled when supported by hardware.

### 3.2 TRAINING CONFIGURATION

Training was conducted using `Unsloth's SFTTrainer` and `FastVisionModel API`, with gradient accumulation enabled to simulate a larger batch size. A cosine learning rate schedule and AdamW optimizer were used. The full configuration is summarized below:

- **Batch size per device:** 2
- **Gradient accumulation:** 4 steps
- **Effective batch size:** 8
- **Epochs:** 2
- **Learning rate:** 2e-4
- **Warmup steps:** 50
- **Optimizer:** adamw\_8bit
- **Weight decay:** 0.01
- **Scheduler:** Cosine decay
- **Max sequence length:** 2048

All models were fine-tuned using the full **RAD-ACE-CoT** dataset. Training was performed on a single **NVIDIA A100 GPU with 40GB VRAM**. While Qwen2.5-VL models (3B and 7B) completed training in approximately **8 hours** each, the larger LLaMA-3.2 Vision-11B model required around **10 hours** to train under the same setup.

To maintain training speed while handling multimodal inputs, we used `UnslothVisionDataCollator`, which efficiently collates visual and textual inputs into tokenized batches.

### 3.3 EXPERIMENT DESIGN

To assess the clinical reasoning capabilities of the fine-tuned models, we curated a test set of **100 high-complexity visual question answering (VQA) samples** sourced from the **VQA-RAD** dataset [23]. These examples were manually selected to emphasize image-text alignment, diagnostic depth, and multi-step medical inference—traits that go beyond simple recognition and test genuine reasoning capacity. This test set was not used during model training, ensuring an objective evaluation of generalization performance.

Instead of relying on traditional lexical overlap metrics (e.g., BLEU [29], ROUGE [30]), which are known to correlate poorly with semantic correctness in medical domains, we adopted the increasingly popular **LLM-as-a-Judge** [25] evaluation methodology. Following best practices from recent multimodal research, we used GPT-4o to assess each model's generated output across five expert-defined criteria:

- **Clinical Relevance:** Is the reasoning focused on medically relevant findings?
- **Factuality / Hallucination:** Are all claims grounded in the image and question? No fabrication?
- **Reasoning Coherence:** Is the reasoning logical, step-by-step, and easy to follow?
- **Completeness:** Are all key aspects of the question and image addressed?
- **Final Answer Quality:** Is the conclusion clinically sound and aligned with the reasoning?

Each response was scored on a 1–5 scale using a structured evaluation prompt, reproduced in A.2. GPT-4o received the image filename, question, ground truth, and model-generated answer, and returned a JSON object with the five scores. This approach enables more fine-grained and medically-aligned assessment than standard automatic metrics.

### 3.4 RESULTS

Table 1: LLM-as-a-Judge evaluation on 100 VQA-RAD samples. Scores range from 1–5 based on GPT-4o judgment across clinical reasoning axes.

Model	Clinical Rel.	Factuality	Reasoning	Completeness	Final Ans.
LLaMA-3.2 Vision-11B	<b>3.16</b>	<b>2.46</b>	<b>3.46</b>	<b>2.46</b>	<b>2.54</b>
Qwen2.5-VL-7B	3.09	2.06	3.09	2.15	2.15
Qwen2.5-VL-3B	2.96	2.00	3.04	2.12	2.12

Sample results and discussion over samples are shared in A.3.

**Overall Performance.** LLaMA-3.2 Vision-11B consistently outperforms the Qwen models across all dimensions, particularly in **reasoning** (3.46) and **clinical relevance** (3.16). Its ability to generate coherent, step-by-step justifications that stay grounded in the clinical prompt likely reflects both the benefit of scale and pretraining alignment.

**Qwen Models: Strong Reasoning, Weaker Closure.** Both Qwen2.5-VL-3B and 7B perform comparably in **reasoning coherence**, scoring above 3.0, which suggests that even mid-sized models are capable of learning the structure of medical reasoning when trained with CoT-style supervision. However, their lower scores in **factuality** and **final answer quality** indicate that generating plausible reasoning does not always translate to correct conclusions. Two main factors likely contribute to this:

- **Lack of Reward Feedback:** These models were trained purely with supervised fine-tuning (SFT). In the absence of task-specific feedback (e.g., through reinforcement learning), the models can learn how to reason but not how to validate their final predictions.
- **Synthetic Data Artifacts:** The training corpus (RAD-ACE-CoT) was constructed from entirely synthetic sources. Inaccuracies or ambiguities in these labels may have introduced hallucination patterns or inconsistencies, particularly affecting factual correctness and closure.

**Size Isn’t Always the Advantage.** Interestingly, the smaller Qwen2.5-VL-3B surpasses the 7B variant—most notably in **factuality** and **final answer quality**. This suggests that a larger parameter count does not always lead to better performance, especially when the training data is noisy or limited. Smaller models may benefit from reduced overfitting and better generalization to cleaner patterns in the data.

**Shared Strength in Structured Reasoning.** A consistent pattern across all models is their relatively higher performance on the **reasoning** axis. This indicates that training on CoT-style prompts—regardless of model size—encourages structured, multi-step explanations. It also highlights a key strength of CoT supervision: improving logical consistency, even in models with modest scale.

## 4 CHALLENGES AND DISCUSSIONS

During the implementation of RAD-ACE, we faced several challenges that shaped the outcomes and highlighted areas for future work.

**1. Synthetic Data and Potential Bias.** Our dataset relies on GPT-generated captions from PubMedVision, which we then processed through DeepSeek-R1 to create CoT-style reports. This pipeline made it easy to scale, but since both stages are synthetic, the data may not fully reflect real-world clinical reasoning. The outputs often looked fluent but might include artifacts that don't generalize well. In future work, we plan to incorporate expert-written reports or clinical sources like PACS data [31] to reduce this bias.

**2. Format Memorization.** One of the most noticeable issues was that the models tended to generate outputs in the same <think>–<answer> format, no matter how the prompt was phrased. This suggests that the model learned to imitate the format rather than generalize instruction-following behavior. We believe adding more instruction-following data from diverse domains like math and coding could improve this, as shown in recent research like X-REASONER [32], where reasoning skills in one domain improved performance in others.

**3. Supervised Fine-Tuning Limits.** We only used supervised fine-tuning (SFT), which helped align model outputs with target formats but didn't offer any feedback on correctness. In tasks like VQA or diagnosis, this becomes a problem. Reinforcement learning (RL) with reward signals (e.g., correct disease prediction) could help the model improve beyond just mimicking patterns.

**4. Evaluation Difficulty.** Because our models generate long CoT-style answers, typical metrics like BLEU or ROUGE weren't useful. Instead, we used GPT-4o as an evaluator, scoring responses on medical relevance, hallucination, and reasoning. This worked reasonably well, but still depends on another LLM's internal biases. The ideal solution would be involving clinical experts in evaluation, especially for real-world deployment.

**5. Engineering Bottlenecks.** Training large VLMs with limited hardware was a constant challenge. Even with Unislot and LoRA optimizations, we hit memory limits, especially with LLaMA-3.2 Vision. We had to experiment with image resolutions, batch sizes, and 8-bit optimizers to get things running. Debugging multimodal pipelines also took time due to inconsistent input expectations across models.

In short, the project highlighted the potential of adapting open-source VLMs for radiology, but also exposed key limitations in data quality, generalization, and evaluation. These will be our main focus areas in the next stages of RAD-ACE.

## 5 CONCLUSION

In this work, we introduced **RAD-ACE**, a framework for training and evaluating vision-language models on clinical reasoning tasks using chain-of-thought supervision. By combining synthetic medical images from PubMedVision with CoT-style outputs generated by DeepSeek-R1, we created a scalable dataset (RAD-ACE-CoT) tailored for radiology-focused report generation. We fine-tuned three open-source vision-language models—Qwen2.5-VL-3B, Qwen2.5-VL-7B, and LLaMA-3.2 Vision-11B—using supervised fine-tuning on this dataset. Evaluation was conducted on a manually curated, high-complexity VQA-RAD test set using the LLM-as-a-Judge methodology with GPT-4o, enabling assessment across five clinical reasoning axes. Our results show that all models, including smaller ones, can learn structured reasoning patterns. However, factual accuracy and answer quality remain key limitations, particularly in models trained solely with SFT and synthetic data.

The study highlights both the promise and the challenges of adapting open-source VLMs to medical domains. Future work will focus on improving final answer quality via reinforcement learning (e.g., GRPO), reducing dataset noise through better synthetic filtering or expert annotation, and incorporating domain-specific evaluation involving clinical professionals. RAD-ACE provides a practical starting point for further exploration into multimodal clinical reasoning at scale.

## REFERENCES

- [1] Radiological Society of North America. (2022, April 7). Radiology preparedness in the wake of conflict. RSNA. <https://www.rsna.org/news/2022/april/radiology-preparedness>
- [2] FreedomIntelligence. (2024). PubMedVision [Dataset]. Hugging Face. <https://huggingface.co/datasets/FreedomIntelligence/Pub>
- [3] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023). The dawn of LMMs: Preliminary explorations with GPT-4V(ision). arXiv. <https://arxiv.org/abs/2309.17421>
- [4] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Zhang, Z. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. <https://arxiv.org/abs/2501.12948>
- [5] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., ... Lin, J. (2025). Qwen2.5-VL Technical Report. arXiv. <https://arxiv.org/abs/2502.13923>
- [6] QwenLM. (2025). Qwen2.5-VL [Computer software]. GitHub. <https://github.com/QwenLM/Qwen2.5-VL>
- [7] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., ... Ma, Z. (2024). The Llama 3 herd of models. arXiv. <https://arxiv.org/abs/2407.21783>
- [8] Meta. (2024). Llama 3.2-11B-Vision [Model]. Hugging Face. <https://huggingface.co/meta-llama/Llama-3.2-11B-Vision>
- [9] Dash, S. K. (2024, October 2). Meta Llama 3.2: A deep dive into vision capabilities. Composio. <https://composio.dev/blog/meta-llama-3-2-a-deep-dive-into-vision-capabilities/>
- [10] DAMO Academy. (n.d.). About us. Alibaba DAMO Academy. <https://damo.alibaba.com/about?language=en>
- [11] Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., ... Qiu, Z. (2025). Qwen2.5 Technical Report. arXiv. <https://arxiv.org/abs/2412.15115>
- [12] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv. <https://arxiv.org/abs/2308.12966>
- [13] Fu, Z., Song, W., Wang, Y., Wu, X., Zheng, Y., Zhang, Y., Xu, D., Wei, X., Xu, T., & Zhao, X. (2025). Sliding window attention training for efficient large language models. arXiv. <https://arxiv.org/abs/2502.18845>
- [14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. arXiv. <https://arxiv.org/abs/1706.03762>
- [15] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., & Zhang, Z. (2024). Qwen2VL: Enhancing vision-language model's perception of the world at any resolution. arXiv. <https://arxiv.org/abs/2409.12191>
- [16] Liu, Z., Guo, L., Tang, Y., Yue, T., Cai, J., Ma, K., Liu, Q., Chen, X., & Liu, J. (2025). VRoPE: Rotary position embedding for video large language models. arXiv. <https://arxiv.org/abs/2502.11664>
- [17] Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). RoFormer: Enhanced transformer with rotary position embedding. arXiv. <https://arxiv.org/abs/2104.09864>
- [18] Shazeer, N. (2020). GLU variants improve Transformer. arXiv. <https://arxiv.org/abs/2002.05202>
- [19] Meta AI. (2024, April 18). Meta Llama 3: The next generation of our open source large language model. Meta. <https://ai.meta.com/blog/meta-llama-3-1/>

- [20] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision
- [21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. <https://arxiv.org/abs/2010.11929>
- [22] Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., & Lin, J. (2025). Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv. <https://arxiv.org/abs/2505.06708>
- [23] Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images [Dataset]. Hugging Face. <https://huggingface.co/datasets/flaviagiammarino/vqa-rad>
- [24] Fan, Z., Liang, C., Wu, C., Zhang, Y., Wang, Y., & Xie, W. (2025). ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification.
- [25] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024). LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods.
- [26] OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., ... Malkov, Y. (2024). GPT-4o System Card. arXiv. <https://arxiv.org/abs/2410.21276>
- [27] Unslloth. (n.d.). Unslloth AI: Open source fine-tuning for LLMs. <https://unslloth.ai/>
- [28] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv. <https://arxiv.org/abs/2106.09685>
- [29] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics. <https://doi.org/10.31115/1073083.1073135>
- [30] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74–81). Association for Computational Linguistics. <https://aclanthology.org/W04-1013/>
- [31] Yu, S., Wu, P., Liang, P. P., Salakhutdinov, R., & Morency, L.-P. (2022). PACS: A dataset for physical audiovisual commonsense reasoning. arXiv. <https://arxiv.org/abs/2203.11130>
- [32] Liu, Q., Zhang, S., Qin, G., Ossowski, T., Gu, Y., Jin, Y., Kiblawi, S., Preston, S., Wei, M., Vozila, P., Naumann, T., & Poon, H. (2025). X-Reasoner: Towards generalizable reasoning across modalities and domains. arXiv. <https://arxiv.org/abs/2505.03981>

## A APPENDIX

### A.1 DATA GENERATION PROMPT

**Role:** You are an Expert Radiologist trained in interpreting CT and MRI images across neuro, thoracic, abdominal, and musculoskeletal domains. You analyze advanced imaging studies using structured visual reasoning and produce evidence-based, formal radiological interpretations.

**Instructions:**

- Assume the image description represents what you see directly with your own eyes.
- Do not mention or imply that you are reading a caption or text.
- Only use the information contained in the visual description.
- Never hallucinate findings or infer beyond what is visually described.
- Interpret all spatial, numerical, or anatomical details confidently and naturally.
- Express detailed reasoning using `<think>` tags.
- Present your final conclusion or diagnosis using `<answer>` tags.
- Use a formal, medically accurate tone throughout.

**Output Format:**

```
<think>

[Step-by-step visual interpretation and clinical reasoning]

</think>

<answer>

[Final diagnosis, interpretation, or conclusion]

</answer>

Image Description: {answer}

Question: {question}
```

Figure 5: Chain-of-Thought Generation Prompt Given to DeepSeek-R1

## A.2 EVALUATION PROMPT

**Role:** You are a medical expert trained in radiology and LLM evaluation. Your task is to evaluate a model-generated output for a medical visual question answering (VQA) task.

**Instructions:**

- Carefully read the image filename, question, model output, and ground truth answer.
- Assess the quality of the model's response based on structured clinical reasoning.
- Score each criterion strictly based on medical and logical merit.
- Return **only** the JSON object. Do **not** include any explanations, markdown, or commentary.

**Scoring Criteria:**

- **1. Clinical Relevance:** Is the reasoning focused on medically relevant findings?
- **2. Factuality / Hallucination:** Are all claims grounded in the image and question? No fabrication?
- **3. Reasoning Coherence:** Is the reasoning logical, step-by-step, and easy to follow?
- **4. Completeness:** Are all key aspects of the question and image addressed?
- **5. Final Answer Quality:** Is the conclusion clinically sound and aligned with the reasoning?

**Output Format:**

```
{  
    "clinical_relevance": 1-5,  
    "factuality": 1-5,  
    "reasoning_coherence": 1-5,  
    "completeness": 1-5,  
    "final_answer_quality": 1-5  
}
```

Figure 6: Evaluation Prompt Given to GPT-4o for Structured Model Assessment

## A.3 SAMPLE RESULTS

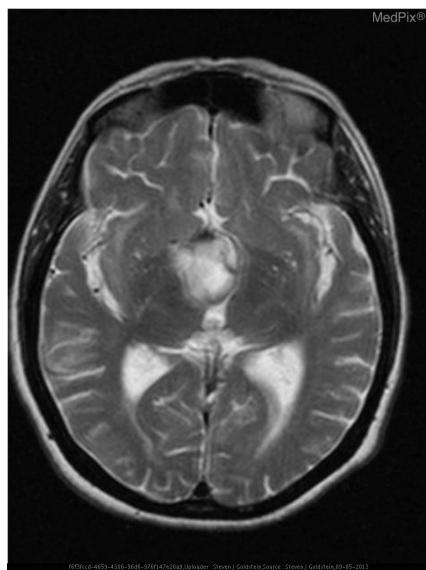
**LLaMA3.2-Vision 11B - Good Result (5 Score Final Answer Quality)**

Figure 7: Example from VQA-RAD

**Question:**

Does the mass affect neighboring structure?

**Ground Truth:**

Yes

**Chain of Thought:**

<think> Okay, so I'm looking at this **MRI image of the brain**. It's an **axial view**, which means it's taken from below, slicing through the brain horizontally. The first thing I notice is that there's a mass in the center of the brain. It looks like it's in the thalamus, which is a deep structure near the center of the brain. The thalamus is important for relaying sensory and motor signals, so any mass there could be significant.

The mass itself is described as heterogeneous. That means it's not uniform—there are different tissue densities within it. On MRI images, this usually shows up as areas of varying brightness. In T2-weighted images, which are good for showing water content, the mass has mixed signal intensities. Some parts are brighter (hyperintense) and others are darker (hypointense). This mix suggests that the mass isn't just one type of tissue; it might have a mix of solid components, cysts, or maybe even areas of necrosis.

I also notice that the mass is causing some distortion in the surrounding brain tissue. The ventricles, which are the fluid-filled spaces in the brain, seem a bit pushed out of their usual shape. This displacement could be due to the mass taking up space and putting pressure on nearby structures. It might also indicate that the mass is growing and affecting the normal anatomy of the brain.

Looking at the adjacent areas, there's some edema or swelling. Edema appears as bright signals on T2-weighted images because it's fluid accumulation. This swelling could be a reaction to the mass, possibly due to inflammation or increased blood flow around the lesion. It might also mean that the mass is causing some damage to the surrounding tissues.

The rest of the brain looks normal in this image. The cerebral hemispheres, basal ganglia, and other structures appear as they should without any obvious abnormalities. This is important because it helps narrow down the possible causes—since everything else seems fine, the issue is likely localized to the thalamic mass.

Putting this all together, the main findings are a heterogeneous mass in the thalamus with mixed signal intensities on MRI, some displacement of surrounding structures, and adjacent edema.

These features could be indicative of a neoplastic process, either benign or malignant, though further imaging or biopsy would be needed for a definitive diagnosis. </think>  
 The axial MRI image reveals a heterogeneous mass in the thalamus with mixed signal intensities on T2-weighted imaging, indicating varied tissue composition. The mass causes displacement of surrounding structures and is associated with adjacent edema. These findings suggest a possible neoplastic process, though further diagnostic workup is necessary for confirmation.  
 <answer> The image shows a **heterogeneous mass in the thalamus with mixed signal intensities on T2-weighted MRI**, causing **displacement of surrounding structures and adjacent edema**, suggesting a **potential neoplastic process**. </answer>

Why is this a good example?

This example (Figure 7) demonstrates strong performance across all reasoning dimensions. The model accurately identifies the **thalamic mass**, details its **heterogeneous signal characteristics**, and clearly states the **displacement of nearby structures and presence of edema**. Most importantly, it directly addresses the question about **impact on neighboring structures**, aligns with the ground truth, and supports its answer with radiologically appropriate observations. This is why it received a perfect score (5) for final answer quality.

#### **Qwen2.5-VL 7B - Poor Result (3 Score Final Answer Quality)**



Figure 8: Example from VQA-RAD

**Question:**

Is the liver normal?

**Ground Truth:**

Yes

**Chain of Thought:**

<think> Okay, so I'm looking at this CT scan of the abdomen and pelvis with contrast. The first thing that catches my eye is something in the right kidney area. It's described as a large mass or lesion there. Since it's contrast-enhanced, I can see how it takes up the dye—maybe it's enhancing, which could mean it's vascularized.

The size is pretty significant: 10 cm by 8 cm. That makes me think it might be a tumor because benign lesions are usually smaller. The fact that it's well-defined suggests it has clear borders, but I shouldn't jump to conclusions without more info.

I notice some internal septations within the mass. Septations can indicate different things—like necrosis or cystic areas. If they're thin and numerous, maybe it's a benign cyst, but thickened walls could suggest malignancy. The enhancement pattern is heterogeneous, meaning parts take

up the contrast differently. This mix might mean varying tissue densities inside, possibly solid and cystic components.

**Looking around the mass, the surrounding structures seem normal. No obvious invasion into nearby organs like the liver or spleen**, which is good because that would complicate things. But I should check for any signs of spread or lymphadenopathy, though those aren't mentioned here.

Putting it all together, this looks suspicious for a malignant tumor. The size and internal characteristics make me lean towards something like renal cell carcinoma, especially since it's in the kidney. However, other possibilities could include metastasis or a large abscess, but the lack of mention of infection signs leans away from that.

So, my conclusion is that this is likely a primary malignant neoplasm of the right kidney.

</think>

<answer> the image depicts a large, well-circumscribed mass in the right kidney, measuring approximately 10 cm x 8 cm, with internal septations and heterogeneous enhancement.

These features raise suspicion for a primary malignant neoplasm, such as renal cell carcinoma.

</answer>

### Why is this a poor example?

In this example (Figure 8), the model completely **ignores the question**, which asks about the liver's condition. Instead, it elaborates on a **mass in the right kidney**, which was not referenced in the prompt. Even though the reasoning is medically plausible, it **fails to assess the liver**—the target of the question—leading to an incorrect focus and a hallucinated response. This mismatch explains the low final answer quality score of 3.