JJ-VMed: A Framework for Automated Concepts, Captions and Explainability of Medical Image

Notebook for the ImageCLEFmed Caption Lab at CLEF 2025

Johanna Angulo[1,*], Jenny Aguilar[1]

[1]School of Science, Engineering and Design, Universidad Europea de Valencia, Paseo de la Alameda, 7, 46010 Valencia, Spain

Abstract

This paper presents JJ-VMed, an experimental multimodal computational framework developed by Jaimage addressing three distinct tasks within the ImageCLEFmedical Caption challenge: concept detection, caption generation, and explainability analysis for medical imaging. The system implements separate but interconnected methodological approaches tailored to each task's specific requirements.

For concept detection and caption generation, the framework employs a systematic four-phase approach utilizing a fine-tuned LLaVA-LLaMA 3 8B model. The methodology incorporates Spanish prompting to enhance multilingual capabilities and cross-linguistic robustness, followed by comprehensive preprocessing, LoRA-based fine-tuning, and systematic post-processing validation. An additional LLaVA-Mistral 7B model was developed with English prompts to address identified limitations, though temporal constraints prevented its full deployment.

The explainability task implements a distinct multi-stage pipeline designed to provide visual grounding for AI-generated medical content. This experimental approach utilizes the concepts and captions from the LLaVA-LLaMA 3 8B model as foundational input, with LLaMA 3.1 merging these outputs to generate separate textual explanations that provide additional context. The pipeline subsequently employs GPT-4o and GTP-4.1 APIs for spatial coordinate mapping, attempting to establish connections between textual explanations and visual features. The final stage implements the Segment Anything Model (SAM) for generating segmentation masks, supplemented by heatmap-based confidence scoring and computer vision techniques including keypoint detection.

The framework generates medical image analyses featuring visual evidence intended to support the generated explanations, constituting a post-hoc approach to medical AI interpretability. This exploratory methodology represents an attempt to contribute to the ongoing research addressing explainable artificial intelligence (XAI) requirements in the medical domain.

Performance evaluation revealed moderate results across tasks: concept detection achieved an F1 score of 0.3982, caption generation obtained an overall score of 0.3043, while the explainability system demonstrated technical feasibility despite not fully meeting challenge objectives. The methodology illustrates both the potential and limitations of current approaches to medical AI interpretability, highlighting areas requiring continued research and development for clinical implementation.

Keywords

Medical Caption, Multimodal AI, AI Explainability, GenAI, Large Language Models, Computer Vision, ImageCLEFmedical

1. Introduction

JJ-VMed, developed by Jaimage, represents an experimental multimodal system for the ImageCLEF 2025 challenge, specifically addressing the ImageCLEFmedical Caption task and its associated evaluation components [1]. This challenge advances multimodal understanding in the medical domain by requiring participants to detect key medical concepts within images, generate clinically accurate captions, and provide transparent explanations that illuminate the model's diagnostic reasoning processes. Our approach centers on Large Language and Vision Assistant (LLaVA)-LLaMA 3 8B, a vision-language model that serves as the foundational architecture for both caption generation and concept detection

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

*Corresponding author: Johanna Angulo (johanna.angulo@gmail.com)

johanna.angulo@gmail.com (J. Angulo); contact@jaguilarweb.com (J. Aguilar)

https://www.linkedin.com/in/johannaangulo/ (J. Angulo); https://www.jaguilarweb.com/ (J. Aguilar)

0009-0005-6965-0604 (J. Angulo); 0009-0008-0262-7778 (J. Aguilar)

tasks [2]. We complement this core system with a post-hoc explainability framework that integrates GPT-4's multimodal capabilities alongside segmentation models, computer vision techniques and specialized medical natural language processing pipelines. This integrated approach enables the system to generate captions and concept labels for radiology images while attempting to provide visual grounding through image region highlighting that corresponds to generated textual content, thereby offering visual evidence supporting the model's outputs. It is important to clarify that our explainability approach does not employ established XAI methods such as SHAP, LIME, GradCAM [3], or attention maps, which directly analyze model internal states. Instead, we implement an experimental post-hoc verification system that independently attempts to validate and visualize the spatial correspondence between generated text and image content. This methodology represents an exploratory contribution toward enhancing trust in medical AI systems through external verification rather than intrinsic model explanation. This paper presents a systematic examination of our experimental system design and empirical results, encompassing the underlying techniques, fine-tuning methodologies, post-hoc verification strategies, and performance evaluation across each challenge task. We provide critical analysis of these results while acknowledging the limitations of our approach and exploring potential directions for future development in medical image analysis systems.

2. Overview of ImageCLEFmedical Caption 2025

The ImageCLEFmedical Caption 2025 challenge is composed of three interconnected tasks: Concept Detection, Caption Prediction, and Explainability [1, 4]. In the Concept Detection Task, systems identify the presence of relevant medical concepts in an image, effectively predicting a set of UMLS (Unified Medical Language System) concept IDs [5] or terms that describe the image's content. This serves as a foundation for captioning by providing the "building blocks" of the scene. In the Caption Prediction Task, systems generate a coherent textual description of the entire image, ideally incorporating the detected concepts and describing their interplay. The Explainability Task, newly introduced in 2025, requires participants to produce an explanation for the caption on a small subset of images – for example, by highlighting image regions and providing additional textual justification. The explainability component is meant to improve interpretability and trust, allowing medical experts to verify why a caption or concept was predicted.

3. Methodology

The primary objective of our research is to develop a multimodal model that generates concept detection and caption prediction. The secondary objective is developing an explainability pipeline combining Natural Language Processing (NLP) and computer vision techniques as an experimental approach.

JJ-VMed medical imaging pipeline represents a three-phase framework that combines fine-tuned multimodal models with explainability mechanisms to produce concept detection, caption prediction and clinically relevant, spatially-grounded medical image analyses.

3.1. Data

All tasks utilized an extended version of the Radiology Objects in Context (ROCO) Version 2 dataset. As in previous editions, the dataset originates from biomedical articles of the PMC OpenAccess subset. We also used the ROCOv2 dataset from the previous year for fine-tuning [6]. The challenge training set (development data) consists of tens of thousands of radiology images (primarily X-rays, CT scans, MRIs, etc.) collected from the biomedical literature, each paired with a figure caption and a set of manually curated UMLS concept labels. Please note the challenge datasets will be described in more detail in the overview paper [1, 4].

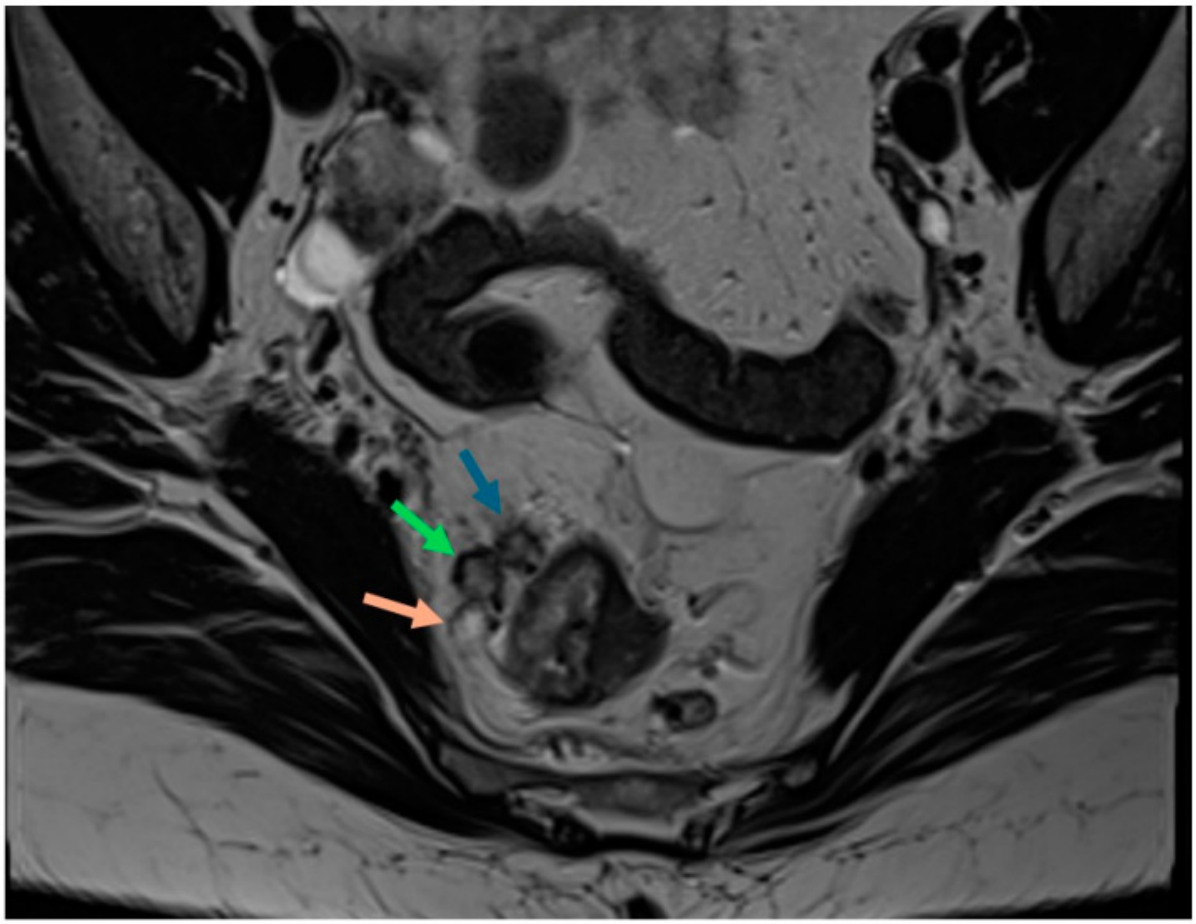

Figure 1: Sample image from the ImageCLEF Challenge test set. Source: ImageCLEFmedical Caption 2025 test 1260, CC BY, Curcean et al., 2024.

3.2. Evaluation Metrics

Each task of the ImageCLEF medical caption challenge uses distinct evaluation metrics [1, 4]. The concept detection task uses F1 scoring methodology [1, 4]. Caption prediction ranking combines six metrics across two evaluation aspects: Relevance Metrics (four metrics) and Factuality Metrics (two metrics). In the explainability task, a human expert radiologist evaluated the quality of each system's generated explanations using a 5-point Likert scale, where 5 represented the highest score. Please refer to the overview paper for more details [1, 4].

3.3. Experimental Setup

For the computational experiments, a combination of two distinct hardware configurations was utilized:

- A desktop workstation equipped with an Intel Core i9 processor and an NVIDIA GeForce RTX 5090 graphics processing unit (GPU) featuring 32 GB of dedicated GDDR7 memory. This machine was employed for all computational tasks, with the exception of operations involving the Segment Anything Model (SAM) due to a technical incompatibility that prevented its use on this specific setup [7, 8].
- A laptop computer, also powered by an Intel Core i9 processor, which included an NVIDIA GeForce RTX 4070 laptop GPU with 8 GB of GDDR6X memory. This laptop was specifically used for Phase III: Computer Vision for Segmentation and Analysis.

The allocation of these resources was based on computational demands and system compatibility. The desktop workstation, with its higher VRAM capacity, was primarily allocated for computationally

intensive training and evaluation phases for all tasks not involving. The laptop, despite its comparatively lower VRAM, specifically handled the computational requirements of the SAM-dependent computer vision tasks. The precise extent of utilization for each GPU varied according to the specific demands of individual experiments.

3.4. Methodology for Concept and Captioning Task

This investigation employs a systematic four-phase approach for medical image concept detection and caption prediction, utilizing fine-tuned multimodal large language models with LoRA (Low-Rank Adaptation). The methodology addresses both technical and clinical requirements through rigorous pre-processing, model adaptation, inference optimization, and validation procedures.

3.4.1. Phase 1: Image Preprocessing

The preprocessing pipeline ensures dataset consistency and model compatibility through systematic image standardization. The process initiates with metadata acquisition from training and validation splits (`train_captions.csv`, `train_concepts.csv`, `valid_captions.csv`, `valid_concepts.csv`), followed by comprehensive image validation and format conversion. Critical preprocessing steps include: (1) RGB format conversion for grayscale and RGBA images, eliminating alpha channels to ensure input consistency; (2) optional resizing to 336×336 pixels using LANCZOS resampling to maintain image quality while standardizing dimensions; (3) JPEG compression [9] at 90% quality for storage optimization. The pipeline implements robust error handling for missing or corrupted files, providing detailed reporting of processing success rates and error classifications. This preprocessing approach was selected to address the heterogeneous nature of medical imaging datasets while maintaining visual fidelity essential for clinical interpretation. The 336×336 resolution represents an optimal balance between computational efficiency and preservation of diagnostic details, as demonstrated in previous vision-language model implementations.

3.4.2. Phase 2: LLaVA-Llama Fine-tuning

The fine-tuning procedure adapts the `llava-llama-3-8b-v1_1-transformers` model for medical imaging tasks using Low-Rank Adaptation (LoRA). This parameter-efficient approach enables domain adaptation while preserving the model's foundational capabilities.

LLaVA (Large Language and Vision Assistant), introduced by Liu et al. in 2023, was designed through visual instruction tuning, enabling it to follow prompts about images in a conversational manner [2]. The LLaVA-LLaMA 3 version represents an updated iteration using a newer LLaMA 3.1 backbone (8B parameters) to improve capability [10]. This multimodal model serves as our base system for automatically generated captions and concepts detection, due to its capability to interpret visual inputs and produce coherent text. The choice of an 8B backbone keeps the model lightweight enough for fine-tuning while still benefiting from LLaMA 3's improvements in language understanding. The fine-tuned model used in this challenge by JJ-VMed is based on a LLaVA model fine-tuned from meta-llama/Meta-Llama-3-8B-Instruct and CLIP-ViT-Large-patch14-336 with ShareGPT4V-PT and InternVL-SFT by XTuner [11, 2].

Model Configuration and Architecture

The base model employs `AutoModelForVision2Seq` with `device_map="auto"` for optimal GPU utilization and bfloat16 precision when supported. Critical architectural modifications include: (1) integration of medical vocabulary through token embedding resizing; (2) `processor patch_size` specification at 14 pixels for vision encoder compatibility; (3) `eos_token` assignment as `pad_token` for consistent sequence handling.

LoRA Implementation

LoRA targets specific transformer components: attention mechanisms (`q_proj`, `k_proj`, `v_proj`, `o_proj`), feed-forward networks (`gate_proj`, `up_proj`, `down_proj`), and multimodal projection layers (`linear_1`, `linear_2`). The configuration employs $r = 16$, $\alpha = 32$, and `dropout` $= 0.05$, balancing adaptation capacity with overfitting prevention. These parameters were selected based on empirical evidence suggesting optimal performance for vision-language tasks with similar model scales.

Training Configuration

Hyperparameter selection follows established practices for medical domain adaptation: `learning_rate` $= 5 \times 10^{-5}$, `batch_size=4` with `gradient_accumulation_steps=8` (effective `batch_size=32`), single epoch training to prevent overfitting on limited medical data. The training employs gradient checkpointing for memory efficiency and evaluation every 250 steps with early stopping based on validation loss.

The optimization strategy utilizes AdamW optimizer with fused implementation for enhanced computational performance. Learning rate scheduling employs cosine annealing with a 3% warmup ratio to ensure stable convergence. Mixed precision training is implemented using bfloat16 precision when supported by hardware, with float16 fallback for compatibility. The maximum sequence length is configured to 1024 tokens to accommodate comprehensive medical descriptions while maintaining computational efficiency. Training monitoring includes performance metrics logged every 50 steps with comprehensive TensorBoard integration for training visualization. The loss computation strategy implements selective masking where input tokens receive label value -100, ensuring only assistant responses contribute to the loss calculation, which is critical for effective instruction tuning. Bias parameters across all targeted modules remain frozen during LoRA adaptation to maintain model stability.

Prompt Engineering Strategy

The training incorporates diverse prompt templates to enhance model generalization: (1) direct image caption requests; (2) medical concept enumeration tasks; (3) combined caption-concept generation; (4) conditional prompts for specific concept queries. This multi-task training approach promotes robust understanding of medical imagery across various clinical scenarios.

The use of Spanish prompts to fine-tune a LLaVA-LLaMA 8B model for concept detection and caption prediction tasks, despite the challenge being in English, is a strategic choice aligned with a broader project to develop a Spanish Multimodal Q&A System. This approach serves as a crucial experimental pipeline within the challenge to cultivate the model's multilingual capabilities. While LLaMA models are primarily English-centric [12], instruction tuning is vital for enhancing their proficiency in other languages [13]. Research indicates that multilingual tuning can be on par with or even surpass monolingual approaches, offering benefits for cross-lingual transfer and robustness [14]. Specifically, by exposing the model to Spanish, we aim to mitigate Image-induced Fidelity Loss (IFL), where LLaVA models often bias responses towards English, particularly after visual input [12]. This bias stems from the language model component and switching to bilingual language model backbones has been shown to reduce IFL [15, 12]. Thus, Spanish prompts improve the model's overall multilinguality [16, 15], making it more appropriate for the Spanish system and potentially enhancing its English performance by fostering a more robust, language-aware internal representation.

Our fine-tuning approach treated the task as a form of instruction following. We constructed a prompt template format for the model that would allow it to learn both concept detection and captioning. During training, each sample was presented to the model randomly and the target output would combine the caption and/or concepts. Here is an example translated into English since we used Spanish prompts.

[System: You are a medical vision-language assistant. Answer with medical terminology when appropriate.]

ID	ImageCLEFmedical_Caption_2025_test_1260,
Caption	"MRI of the pelvis showing a large mass (green arrow) in the right adnexa with a small amount of free fluid (blue arrow) and a small amount of blood (red arrow)."

ID	ImageCLEFmedical_Caption_2025_test_1260,	
CUIs	C0024485	C0030797
	C0042149	C0007874
	C0227844	

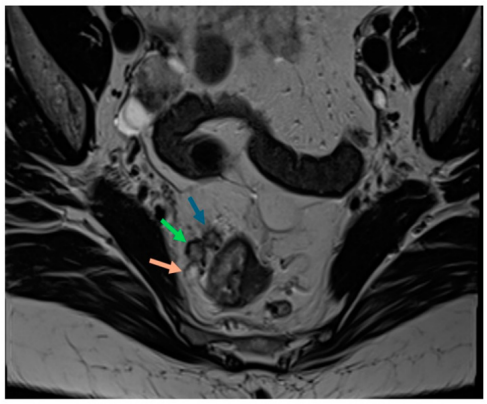

Figure 2: Sample caption and concepts generated by the JJ-VMed System. Source: Elaborated by the authors using ImageCLEFmedical Caption 2025 test 1260, CC BY, Curcean et al., 2024.

User: Describe the following radiology image in detail and list the key findings.

[Image]

Assistant:

Regarding outputs, for the challenge we experimented with two formats:

Concepts only format: The model outputs a list of "Concepts" with its Concept Unique Identifiers (CUIs). This format explicitly trains the model to identify and name the important concepts (after seeing the image) before generating the fluent sentence.

Caption only format: The model outputs just the caption sentence(s), implicitly learning to mention the important concepts within it. This is closer to how radiologists write captions (embedding key terms in the description).

The outputs were two .csv files with two columns:

- Concepts: column ID with the image name and column CUIs with the identified concepts.
- Captions: column ID with the image name and column Caption with the predicted captions.

3.4.3. Phase 3: Model Inference

Inference procedures address dual objectives: concept detection and caption generation, each employing task-specific optimization strategies.

Concept Detection Protocol

Concept inference utilizes structured prompts in Spanish: "Enumera los conceptos médicos clave (CUIs) observados o inferidos en esta imagen." The model generates natural language medical terms subsequently mapped to Concept Unique Identifiers (CUIs) through a comprehensive dictionary matching system. The CUI mapping process implements fuzzy matching algorithms to handle terminology variations and synonyms, ensuring robust concept identification. The `convert_natural_to_cui` function processes comma-separated natural language outputs, yielding semicolon-separated CUI codes while maintaining concept uniqueness and ordering.

Caption Generation Methodology

Two caption generation approaches were implemented: (1) direct description prompting; (2) concept-conditioned generation. The optimized approach incorporates previously identified natural language

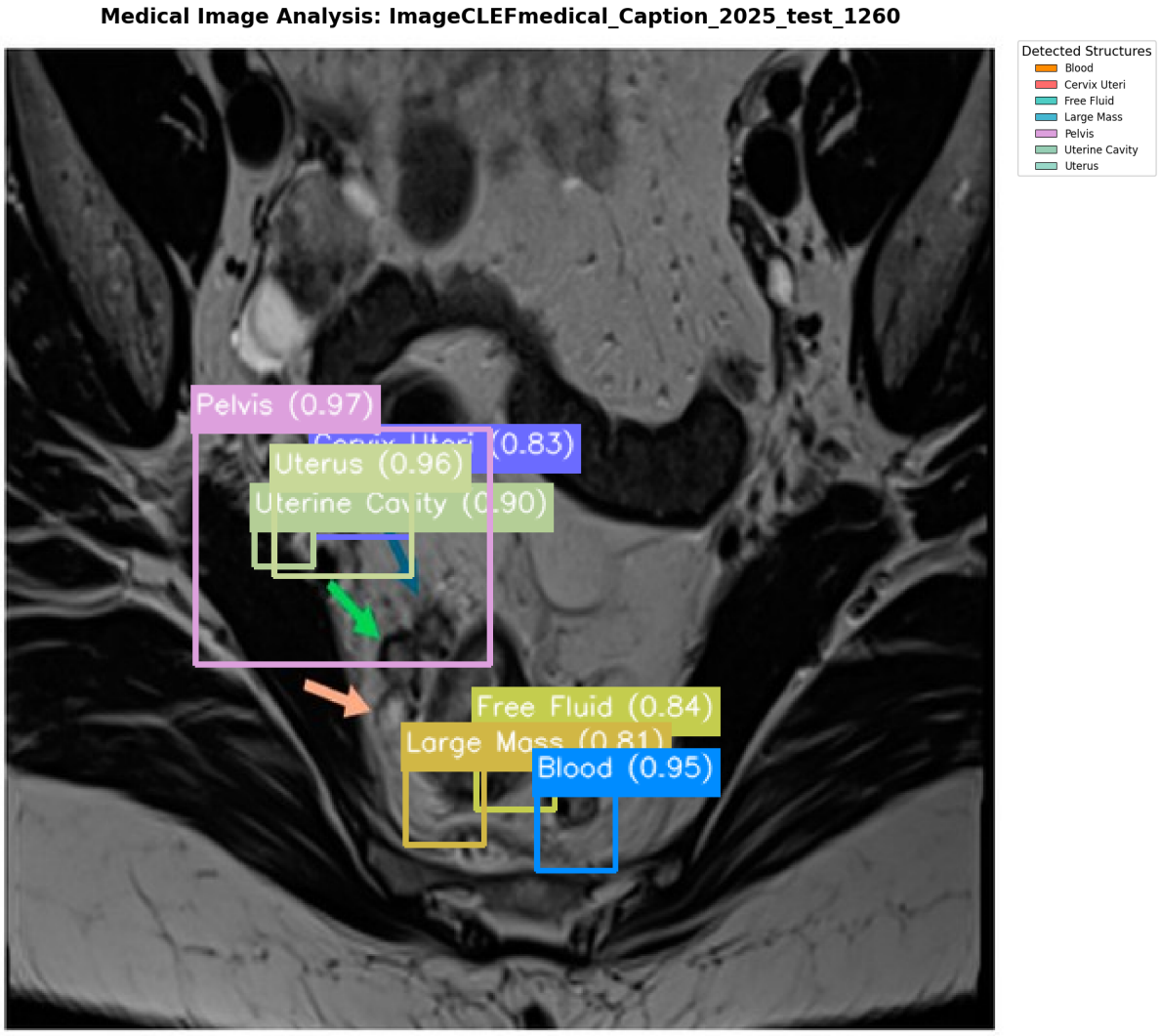

Figure 3: Image with the concepts and captions predicted by our model. Source: Elaborated by the authors using ImageCLEFmedical Caption 2025 test 1260, CC BY, Curcean et al., 2024.

concepts into prompts: "Describe esta imagen médica enfocándose en los siguientes conceptos clave: concepts." This conditioning strategy enhances caption relevance and clinical accuracy by leveraging concept detection outputs. Generation parameters include the `max_new_tokens` configuration and optional sampling strategies (`do_sample`, temperature) to balance creativity and factual accuracy. Batch processing with checkpointing mechanisms ensures computational efficiency and recovery capabilities.

Technical Implementation Details

Model loading employs optional 4-bit quantization using `BitsAndBytesConfig` for memory-constrained environments. PEFT adapter integration allows seamless fine-tuned model deployment while maintaining base model integrity. The inference pipeline implements comprehensive error handling and progress tracking for large-scale dataset processing.

3.4.4. Phase 4: Post-processing and Validation

Post-processing ensures output format compliance and content consistency through systematic validation procedures.

Concept Validation Protocol

The `clean_and_process_cui_string` function addresses common output formatting issues: (1) duplicate CUI removal while preserving first occurrence order; (2) whitespace normalization and empty string elimination; (3) semicolon-separated format standardization. This validation ensures submission compliance with evaluation framework requirements.

Caption Refinement Procedures

Caption cleaning targets generation artifacts through quote character normalization. The `clean_quotes` function removes erroneous triple quotes, double quote sequences, and inconsistent quotation marks. Subsequent CSV formatting ensures proper encapsulation of caption content while maintaining identifier integrity.

3.4.5. Methodological Justification

This four-phase approach addresses specific challenges in medical image analysis: (1) dataset heterogeneity through systematic preprocessing; (2) domain adaptation requirements via LoRA fine-tuning; (3) clinical relevance through concept-conditioned generation; (4) evaluation compliance through rigorous post-processing. The LoRA adaptation strategy was selected over full fine-tuning to prevent catastrophic forgetting while enabling domain-specific learning. The concept-conditioned caption generation approach represents a novel contribution, leveraging detected medical concepts to guide description generation toward clinically relevant content.

3.4.6. Implementation Constraints and Methodological Implications

To augment the conceptual richness and detection granularity of the foundational text generation process, we initiated the development of a secondary fine-tuned LLaVa-Mistral 7B model specifically optimized for enhanced medical concept identification. This model was designed to complement the primary LLaVa Llama 8B model by providing more comprehensive and nuanced detection of medical entities, anatomical structures, and pathological findings within medical images. Despite successful model fine-tuning completion, temporal constraints imposed by the challenge deadline prevented the execution of full inference processes on the target dataset. Consequently, the enhanced concept detections from this secondary model could not be integrated into the final pipeline implementation. This limitation represents a significant methodological constraint that potentially affected the comprehensiveness of concept coverage in the final explainability outputs. The absence of this enhanced model's contributions may have resulted in reduced recall for subtle or complex medical entities that would have benefited from the more sophisticated detection capabilities.

3.5. Methodology for Explainability task

The approach outlines a multi-stage pipeline designed to provide visual explanations for AI-generated medical image captions. Although this system did not fully meet the challenge objective (as discussed in the *Explainability Discussion* subsection in the Analysis section), this experimental endeavour represents an attempt to contribute to the ongoing research addressing explainable artificial intelligence (XAI) requirements in the medical domain, where the inherent opacity of deep learning models can hinder trust and adoption in clinical settings [17, 18].

3.5.1. Introduction and Foundational Data

The foundational data for this project originates from textual content—specifically, initial concepts and captions generated by a fine-tuned LLaVa Llama 8B model. LLaVA (Large Language-and-Vision Assistant) models, like LLaVA-Med and Visual Instruction Tuning, are pre-trained on vast multimodal datasets, making them capable of interpreting and generating text from visual inputs.

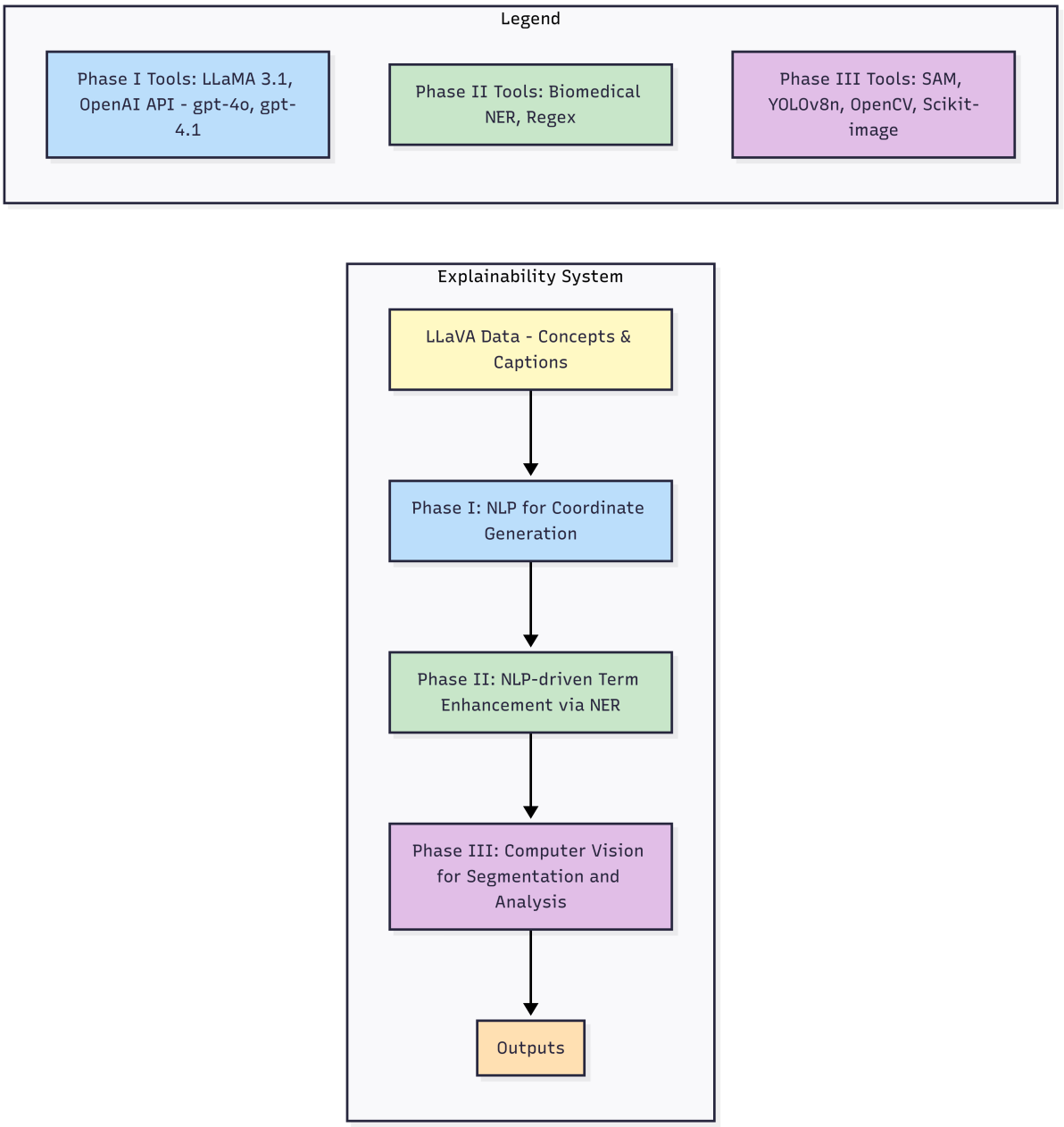

Figure 4: Explainability System Architecture (Source: Developed by the authors).

3.5.2. Phase I: Natural Language Processing for Coordinate Generation

Natural Language Processing begins with the Llama 3.1 model merging concepts and captions into refined textual outputs [10]. Large language models (LLMs) and transformer-based architectures, such as those discussed in the GPT-4 Technical Report [19], have demonstrated good performance in understanding and generating nuanced natural language, proving crucial for synthesizing refined explanations [19]. The emphasis on high-quality, natural language forms the basis for human-interpretable explanations [17]. Subsequently, the initial computational phase is dedicated to translating abstract textual concepts into approximate spatial coordinates, specifically bounding box coordinates (`x`, `y`, `width`, `height`) or arrow-tip locations. This transformation is executed via a pipeline that interfaces with OpenAI API models [19]. For vision-language tasks, the `gpt-4o` model is employed, leveraging its multimodal capabilities to interpret medical images and their associated captions [19]. GPT-4 exhibits human-level performance on professional and academic benchmarks [20, 21] and processes image and text inputs to

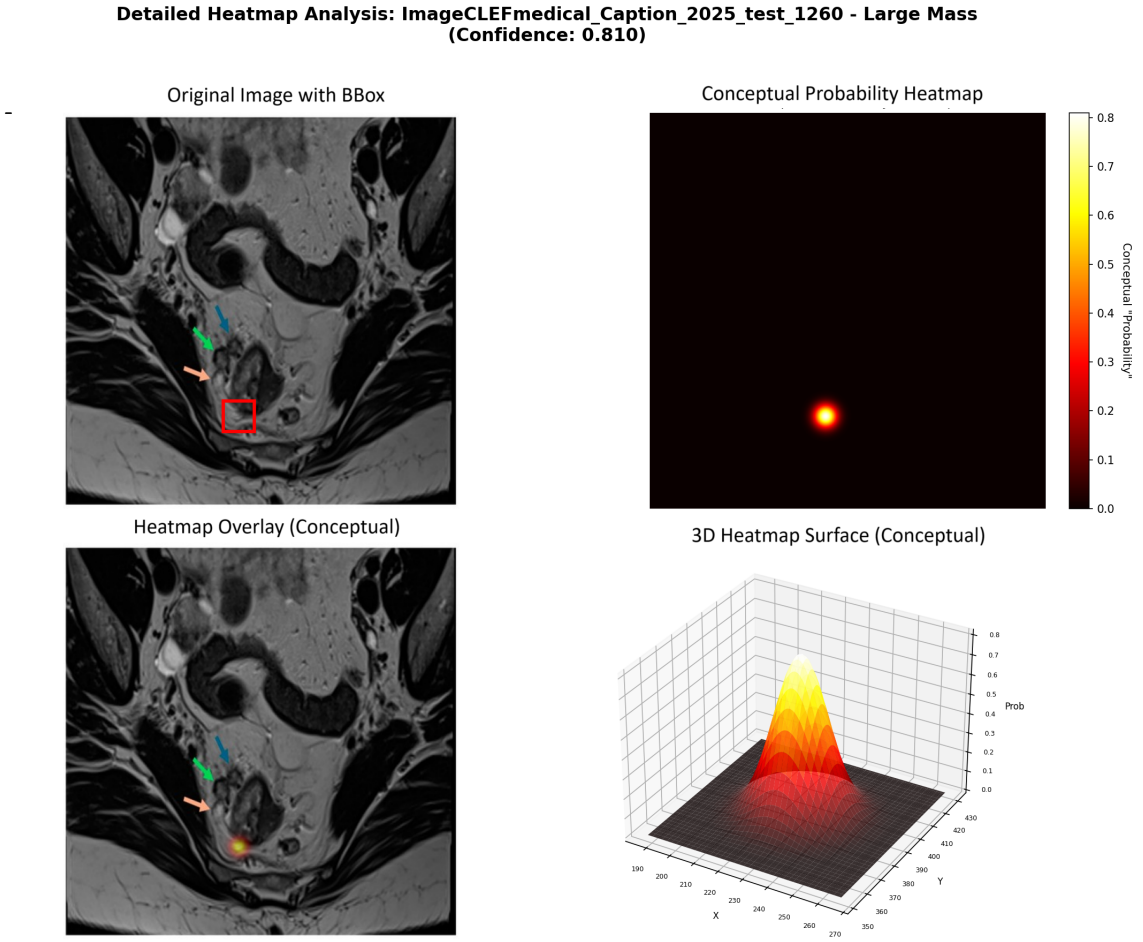

Figure 5: Sample heatmap for image 1260 of the test set. Source: Elaborated by the authors using ImageCLEFmedical Caption 2025 test 1260, CC BY, Curcean et al., 2024.

produce text outputs [19]. Subsequently, for refinement and validation, the `gpt-4.1` model is utilized [19]. The system relies on robust NLP libraries, including `openai` (v1.x), `pandas`, `json`, and `re`, for efficient data handling.

To achieve high precision in term normalization, several NLP techniques are rigorously applied. Rule-based filtering is implemented using regular expressions to exclude generic modality terms like "MRI" or "X-ray". This ensures analysis focuses on anatomically significant entities [18]. A preprocessing step called *group compound terms* uses a dictionary of compound terms (e.g., *small intestine*, *left axillary region*). This maintains token integrity and prevents fragmentation of terms like *right lower lobe*, which is crucial in medical text processing [17].

A two-model prompting strategy underpins the coordinate generation process. First, `gpt-4o` is prompted to analyse the image, caption, and normalized terms, estimating bounding box coordinates or arrow-tip locations. Then, `gpt-4.1` validates and refines this output. If a bounding box is absent but an arrow exists, a 40x40 pixel box is inferred. This yields a curated dataset in `sam_coord.csv` and JSON, ready for segmentation[7].

3.5.3. Phase II: NLP-driven Term Enhancement via Named Entity Recognition (NER)

To augment medical concepts, a dedicated NLP stage extracts terms directly from captions. This employs a hybrid Named Entity Recognition (NER) approach: a biomedical model plus rule-based techniques. The biomedical NER model `d4data/biomedical-ner-all`, accessed via the `transformers` library (v4.x), is specifically designed for clinical entity recognition [22]. Complementing this,

`extract_medical_terms_enhanced` regular expressions capture linguistic patterns (e.g., *1.5 cm hypoechoic mass*, *hematoma*, *stenosis*).

This dual approach ensures comprehensive terminology extraction. Newly identified terms are compared against `sam_coord.csv`. Novel terms re-enter Phase I, generating coordinates that enrich `sam_coord.csv`. This refinement enhances localization of clinical information.

3.5.4. Phase III: Computer Vision for Segmentation and Analysis

This final phase uses the coordinates to segment images and generate analytical outputs. It integrates advanced CV models and libraries tuned for medical imaging [23].

For segmentation, the Segment Anything Model (SAM) is used [8, 24]. We used `sam_vit_h_4b8939.pth` model and `segment-anything` on PyTorch.

For detection, the `YOLOv8n` model from `ultralytics` is employed [25]. YOLO is known for real-time object detection [26], with YOLOv7 and YOLOv8 being recent innovations [25, 26]. Here, `YOLOv8n` enhances recall by detecting overlooked entities [25]. Detections with Intersection over Union (IoU) < 0.5 are added for segmentation. Time constraints prevented a custom anatomical model, so a generic `YOLOv8n-seg.pt` model was used, which introduced noise due to non-medical tags.

Computer Vision Techniques for Masking:

- Direct Bounding Box Prompt: The primary strategy was to prompt SAM with the bounding box generated.
- Keypoint-Based Prompting: If the confidence of the initial mask was low, the system escalated to a more nuanced prompting strategy using local feature descriptors [27]. These were not used for matching, but as a set of salient points to guide SAM's attention within the bounding box. The keypoint detectors used were SIFT (Scale-Invariant Feature Transform), FAST (Features from Accelerated Segment Test) and LoG (Laplacian of Gaussian) [28].

Feature detection uses `opencv-python` (v4.x) and `scikit-image` (v0.x). Outputs include diagnostic plots and annotated images with legends, aligning with the heatmap concept for explainability.

In the Explainability task, three types of outputs were generated:

- Bounding Boxes: Colored rectangular regions that mark areas of interest within the medical images. Each image is accompanied by a legend indicating the specific medical entities identified, with multiple bounding boxes possible per image to highlight different anatomical structures or pathological findings.
- Heatmaps: Comprehensive visualization sets consisting of four representations per identified object or label: the original image with bounding boxes, a conceptual probability heatmap showing confidence distributions, a three-dimensional heatmap surface providing depth visualization, and a heatmap overlay that combines visual attention with the original image content.
- Internal Metrics: Quantitative measurements related to the computer vision techniques employed, providing technical insights into the model's performance and decision-making processes during the explainability generation phase.

CSV exports detail confidence scores and metrics for each strategy. This supports AI explainability and clinical decision-making. The approach is exploratory, targeting transparency and faithfulness—the degree to which explanations reflect model reasoning—and plausibility—how well they align with human understanding [18]. Since XAI methods often fail to meet full clinical requirements, this structured approach aims to bridge that gap and offer interpretable outputs to clinicians [18].

Table 1
Concept Detection Results. These are the official results on the independent test set for the ImageCLEFmedical Caption 2025 task.

Method	F1	F1 secondary
Our System	0.3982	0.8329

Table 2
Caption Prediction Results. These are the official results on the independent test set for the ImageCLEFmedical Caption 2025 task.

Similarity	BERTScore	ROUGE-1	BLEURT	Rel. Avg.	UMLS C. F1	AlignS	Fact. Avg.	Overall
0.8251	0.5953	0.2389	0.3094	0.4922	0.1366	0.0964	0.1165	0.3043

Table 3
Explainability Task - Human Evaluation Results. These are the official results on the independent test set for the ImageCLEFmedical Caption 2025 task.

C.R.	Clin.A.	C.LOD	C.f.	M.C.R.	V-T coh.	Com.V	V.foc.	M.V.rating	A.M.	Overall
3.4	2.4	2.8	4.1	3.2	1.9	1.9	1.9	1.9	2.0	2.6

4. Results

This section presents JJ-VMed results under the username Jaimage. The three tasks were evaluated on the official ImageCLEFmedical test sets [1, 4]. We report performance on all tasks in which our systems participated: Concept Detection, Caption Prediction and Explainability. The results include different metrics for each task as explained earlier. For more details refer to the Overview paper [4]. For Concept Detection results see Table 1, for Caption Prediction results see Table 2 and for Explainability task results see Table 3. These are the official results on the independent test set for ImageCLEFmedical Caption 2025 Tasks [4].

5. Analysis and Discussion

5.1. Concept Detection Performance Analysis

5.1.1. Primary Performance Metrics and Limitations

Our system achieved an F1 score of 0.3982 in the concept detection task, representing a substantial performance deficit compared to the leading approach (F1 = 0.5888). This 32.4% performance gap indicates significant limitations in our methodology's ability to accurately identify and classify medical concepts within the target dataset. The moderate F1 score suggests fundamental weaknesses in either our feature extraction processes, concept classification mechanisms, or both.

Critical Analysis of Underperformance: Several methodological factors likely contributed to this suboptimal performance. First, the reliance on a single fine-tuned LLaVA Llama 8B model may have introduced inherent limitations in concept recognition capabilities, particularly for specialized medical terminology. The fine-tuning parameters may not have been optimally configured for the specific medical imaging domain represented in the challenge dataset.

Language-Specific Performance Degradation: A significant contributing factor to the observed performance deficit can be attributed to the use of Spanish prompting in our primary model implementation. Medical terminology translation and cross-linguistic concept mapping introduce additional complexity layers that can adversely affect concept detection accuracy. Spanish medical terminology may not align precisely with the English-based training datasets commonly used in large vision-language

models, potentially causing systematic misclassifications or missed detections.

5.1.2. Secondary Metric Performance and Contextual Analysis

While our approach demonstrated improved performance on the secondary F1 metric (0.8329), this result requires critical contextualization. Although this score appears substantially higher than our primary F1, the 12.2% performance gap compared to the leading team's secondary F1 (0.9484) remains significant and should not be underestimated. This approximately 10% differential in a curated subset of key concepts indicates that our methodology still fails to achieve optimal performance even when evaluated on the most clinically relevant features.

Methodological Implications: The improved secondary F1 score suggests that our approach demonstrates reasonable proficiency in detecting prominent or well-defined medical concepts while struggling with more nuanced, rare, or complex findings. This pattern indicates potential insufficient training exposure to the full spectrum of medical concept diversity present in clinical practice.

5.1.3. Methodological Response to Identified Limitations

During the challenge development phase, we observed suboptimal concept detection performance in our preliminary evaluations, though without access to comparative benchmarks or final official results. Recognizing these potential deficiencies and anticipating the need for enhanced performance, we initiated the fine-tuning of a secondary LLaVA-Mistral 7B model utilizing English prompts to address both the language-specific limitations and enhance overall concept detection granularity. The English prompting strategy was specifically implemented to mitigate the cross-linguistic performance degradation observed in our primary Spanish-prompted model. This secondary model demonstrated enhanced descriptive output quality in preliminary evaluations, providing more detailed concept identification compared to our baseline approach.

The successful development of the LLaVA-Mistral 7B model confirms the technical viability of multi-model ensemble approaches for enhanced concept detection. However, the incomplete deployment due to temporal constraints represents a critical methodological limitation that prevented empirical validation of this enhancement strategy's effectiveness. Future implementations must prioritize earlier integration of secondary models to ensure comprehensive evaluation within project timelines.

5.2. Caption Prediction Performance Analysis

5.2.1. Overall Performance Assessment and Competitive Context

Our caption generation model achieved an overall score of 0.3043, representing approximately 88.7% of the leading team's performance (0.3432). While this result demonstrates moderate competitiveness, the 11.3% performance gap indicates substantial room for improvement in caption generation quality and accuracy.

5.2.2. Strengths in Textual Similarity and Content Overlap

Our model demonstrated notable strengths in specific evaluation dimensions. The high textual similarity score (0.8251) and competitive BERTScore Recall (0.5953 vs. 0.5977 for the leading team) indicate effective content overlap and phrasing alignment with ground-truth reports. These metrics suggest that our approach successfully captures general descriptive patterns and maintains reasonable linguistic coherence in generated captions.

The moderate ROUGE-1 (0.2389) and BLEURT (0.3094) scores reflect acceptable textual overlap and overall caption quality as measured by established metrics. The average relevance score of 0.4922 indicates that our generated captions incorporate a substantial portion of key information from reference reports, demonstrating reasonable content coverage.

5.2.3. Critical Analysis of Domain-Specific Performance Deficits

Medical Terminology and Concept Accuracy: Our system exhibits significant weaknesses in domain-specific medical accuracy, as evidenced by the UMLS concept F1 score of 0.1366, representing a 24.8% deficit compared to the leading approach (0.1816). This substantial gap indicates systematic difficulties in accurately identifying, incorporating, and maintaining consistency with specialized medical terminology and specific clinical findings.

Visual-Textual Alignment Deficiencies: The AlignScore of 0.0964 (versus 0.1375 for optimal performance) reveals a critical limitation in aligning image content with generated descriptions. This 30.1% performance deficit suggests that our model frequently generates descriptions containing details insufficiently supported by visual evidence, indicating potential hallucination tendencies or inadequate visual feature extraction capabilities.

Factual Consistency and Reliability Issues: The factuality average of 0.1165, compared to the leading result of 0.1596, represents a 27.0% performance gap that raises serious concerns about clinical applicability. This deficit indicates recurring factual inconsistencies and possible hallucinations in generated text, which constitute fundamental reliability issues for medical applications where accuracy is paramount.

5.2.4. Systematic Analysis of Performance Limitations

Root Cause Analysis: The observed deficiencies likely stem from several interconnected methodological limitations:

1. **Fine-Tuning Language:** Our fine-tuning process used Spanish prompts and the model although understand Spanish may have struggle to find specialized medical terminology for medical imaging contexts, resulting in generic rather than clinically precise descriptions.
2. **Visual Feature Extraction Limitations:** The visual encoding components may lack sufficient granularity to capture subtle medical imaging features necessary for accurate clinical description.
3. **Cross-Modal Integration Deficiencies:** The alignment between visual and textual representations appears suboptimal, leading to descriptions that fail to accurately reflect image content.
4. **Limited Medical Knowledge Integration:** The absence of explicit medical knowledge bases or fact-checking mechanisms likely contributes to factual inaccuracies and terminology misuse.

5.2.5. Methodological Reflection

This performance analysis reveals that our experimental approach, while demonstrating technical feasibility, requires substantial refinement to achieve clinical relevance. The systematic underperformance across multiple evaluation metrics indicates fundamental limitations in our current methodology that extend beyond simple parameter optimization. The language-specific performance degradation and incomplete deployment of enhancement strategies highlight the importance of comprehensive planning and early implementation of methodological improvements in future research endeavors.

5.3. Explainability Discussion

The presented framework is classified as a post-hoc, concept-grounding explainability system. Its fundamental objective is to provide clinical interpretability by independently verifying the model's textual outputs against the visual evidence present in the image.

5.3.1. Acknowledging the Deviation from Direct Black-Box Explanation

It is important to acknowledge that the primary objective of this framework is not to elucidate the internal computational pathways of the "black-box" captioning model. This approach stands in stark contrast to intrinsic or ante-hoc methods, which are designed to be transparent by nature, and other post-hoc methods, such as Grad-CAM [17], which aim to explain a model's decision-making process by

inspecting its internal state (e.g., gradients, activations) [29]. While explainable AI (XAI) broadly aims to make AI models more transparent, interpretable, and understandable, critics of post-hoc explanations correctly assert that they merely approximate, rather than replicate, the actual reasoning processes of black-box systems. A truly "complete" post-hoc explanation would, in essence, equate to the original model itself, thus negating the need for the original mode [18].

5.3.2. Justification of the Post-Hoc Approach for Clinical Interpretability

Despite this acknowledged deviation from direct black-box explanation, the adoption of a post-hoc explainability framework was necessitated by pragmatic constraints encountered during the challenge, including the unavailability of a fully developed intrinsic explainability module and demanding challenge timelines [29, 30]. The resultant strategy was to prioritize a method that could effectively enhance a clinician's trust in the final output, even if the model's internal logic remained opaque [30, 18, 29].

This approach is firmly rooted in the concept central to human-centric XAI, positing that for clinical adoption, trust can be fostered by demonstrating that a model's conclusions are factually correct and visually verifiable [30, 18]. The system attempts to achieve this by employing an independent pipeline to answer a crucial question for clinicians: "Given the model's claim, is there corroborating evidence in the image?".

Though indirect, this external audit approach offers several distinct benefits:

- **Modularity:** It can be applied to any captioning model, underscoring its versatility.
- **Integration of Specialized Models:** It allows for the integration of task-specific models, such as SAM (Segment Anything Model), optimized for segmentation verification [8].
- **Alternative Visualization Approach:** It enables the generation of segmentation masks through SAM [8], offering an alternative to heatmap-based explanations. This approach provides spatially defined regions that may assist in clinical interpretation, though empirical validation of radiologist preferences between visualization methods remains to be established.

Furthermore, defenses of post-hoc explanations highlight their utility even without replicating internal model reasoning. They can improve users' functional understanding of black-box systems, increase the accuracy of clinician-AI teams (e.g., radiologist-AI teams have shown improved accuracy with saliency masks), and assist clinicians in justifying their AI-informed decisions. Such explanations empower users to better discriminate between correct and incorrect outputs [31].

5.3.3. Limitations and Complementary Nature

Nevertheless, it is critical to acknowledge the fundamental limitation of this external audit approach: its inability to diagnose why a model fails. It can effectively detect a hallucinated finding but cannot explain its origin within the source model's architecture. Therefore, this work is best viewed as a pragmatic exploration into building trust via external, multimodal verification. It complements, rather than replaces, the critical role of intrinsic explainability methods. For instance, techniques like Grad-CAM are invaluable for model debugging and understanding feature attribution by localizing where in the image the model "looked" [17]. This differs from the presented system's output of grounded segmentation masks. To illustrate this distinction, Appendix B provides a comparative Grad-CAM visualization (obtained post-challenge from an updated version of the system). This visualization showcases how an intrinsic method highlights the captioning model's regions of interest, an approach that differs from our challenge system's output of grounded segmentation masks.

Thus, while this post-hoc verification serves a vital role in fostering clinical trust and providing clear, verifiable evidence, it is part of a broader explainable AI ecosystem necessary for a holistic understanding of AI system behavior and for the comprehensive concept of trustworthy AI, which extends beyond mere explainability to include dimensions such as fairness, safety, privacy, accountability, and robustness.

6. Future Work

Based on our comprehensive evaluation and analysis of the JJ-VMed system's performance across the ImageCLEF medical captioning tasks, we have identified several critical areas for improvement and promising research directions. Our current results, reveal specific limitations that provide clear pathways for future enhancement.

Several research directions warrant exploration for improving our system in future editions. Some of them are under development.

6.1. Methodological Improvements and Future Directions

Our results indicate that while our approach achieves reasonable general language quality and content relevance, significant improvements are required in medical domain specificity and factual accuracy. Future developments should prioritize:

1. **Enhanced Medical Knowledge Integration:** Incorporating structured medical knowledge bases and terminology validation systems
2. **Improved Visual Feature Extraction:** Implementing medical imaging-specific feature extraction methods
3. **Robust Fact-Checking Mechanisms:** Developing explicit verification systems for generated medical content
4. **Cross-Modal Alignment Optimization:** Enhancing the integration between visual and textual model components

6.2. Capturing Clinical Content

Integrate more multimodal vision-language foundation models to better capture clinical content [32]. Ensembling complementary image encoders can produce more fluent and contextually accurate descriptions [33]. By adopting such architectures – powerful vision backbones and generative transformers – captioning systems can better describe complex scenes and rare pathologies in natural, expert-like language. This enhancement directly addresses our current limitations in UMLS concept detection (F1: 0.1366) and factual alignment, potentially improving both clinical accuracy and terminology precision in generated captions.

6.3. Hybrid Transformers + CNN Approach

The concept detection component of our system presents opportunities for significant improvement through architectural innovation. While Transformer-based models have demonstrated competitive performance in medical image analysis, they have yet to fully surpass Convolutional Neural Networks (CNN)-based approaches in all medical imaging contexts. However, these models have shown good results at capturing global contextual relationships, which are crucial for comprehensive medical image interpretation [34]. We propose developing a hybrid architecture that strategically combines the strengths of both paradigms: leveraging Transformers' superior global context modeling capabilities alongside CNNs' proven effectiveness in local feature extraction and spatial relationship detection [23]. This integrated approach could potentially overcome the individual limitations of each architecture while maximizing their complementary strengths.

6.4. Evaluation and Validation Framework

Future research will also focus on developing more comprehensive evaluation methodologies that better capture the nuanced requirements of medical image captioning systems. This includes establishing stronger correlations between automated metrics and clinical utility assessments, potentially through expanded human expert evaluation protocols and task-specific evaluation criteria.

These improvements, grounded in our current performance analysis, ongoing research efforts, and emerging research trends, provide a clear roadmap for advancing the JJ-VMed system toward a more robust, clinically accurate, and explainable medical image captioning system.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Claude in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEF medical 2025 – medical concept detection and interpretable caption generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[2] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, 2024, pp. 26286–26296. doi:10.1109/cvpr52733.2024.02484.

[3] R. Rs, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science (LNCS), Madrid, Spain, 2025.

[5] L. Amos, D. Anderson, S. Brody, A. Ripple, B. L. Humphreys, UMLS users and uses: a current overview, Journal of the American Medical Informatics Association 27 (2020) 1606–1611. URL: https://doi.org/10.1093/jamia/ocaa084. doi:10.1093/jamia/ocaa084.

[6] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[7] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, Y. Jin, Medical SAM adapter: Adapting segment anything model for medical image segmentation, Medical Image Analysis 102 (2025) 103547. doi:10.1016/j.media.2025.103547.

[8] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, Y. Zhang, Segment anything model for medical image analysis: An experimental study, Medical Image Analysis 89 (2023) 102918. URL: https://www.sciencedirect.com/science/article/pii/S1361841523001780. doi:https://doi.org/10.1016/j.media.2023.102918.

[9] M. Rabbani, R. Joshi, An overview of the JPEG 2000 still image compression standard, Signal Processing: Image Communication 17 (2002) 3–48. URL: https://www.sciencedirect.com/science/article/pii/S0923596501000248. doi:https://doi.org/10.1016/S0923-5965(01)00024-8, JPEG 2000.

[10] Meta, Meta Llama 3.1, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024.

[11] X. Contributors, XTuner: A toolkit for efficiently fine-tuning LLM, https://github.com/InternLM/xtuner, 2023.

[12] M. Hinck, C. Holtermann, M. Olson, F. Schneider, S. Yu, A. Bhiwandiwalla, A. Lauscher, S.-Y. Tseng, V. Lal, Why do LLaVA vision-language models reply to images in english? (2024) 13402–13421. doi:10.18653/v1/2024.findings-emnlp.783.

[13] D. Aftab, S. Davy, Tailored-LLaMA: Optimizing few-shot learning in pruned LLaMA models with task-specific prompts, in: U. Endriss, F. S. Melo, K. Bach, A. Bugarin-Diz, J. M. Alonso-Moral, S. Barro, F. Heintz (Eds.), ECAI 2024 - 27th European Conference on Artificial Intelligence, Including 13th Conference on Prestigious Applications of Intelligent Systems, PAIS 2024, Proceedings, IOS Press BV, Santiago de Compostela, Spain, 2024, pp. 3844–3850. URL: https://researchprofiles.tudublin.ie/en/

publications/tailored-llama-optimizing-few-shot-learning-in-pruned-llama-model. doi:10.3233/FAIA240642.

[14] D. Stap, E. Hasler, B. Byrne, C. Monz, K. Tran, The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities, https://www.amazon.science/publications/the-fine-tuning-paradox-boosting-translation-quality-without-sacrificing-llm-abilities, 2024. Amazon Science.

[15] D. Zhu, P. Chen, M. Zhang, B. Haddow, X. Shen, D. Klakow, Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 388–409. URL: https://aclanthology.org/2024.emnlp-main.24/. doi:10.18653/v1/2024.emnlp-main.24.

[16] P. Chen, S. Ji, N. Bogoychev, A. Kutuzov, B. Haddow, K. Heafield, Monolingual or multilingual instruction tuning: Which makes a better alpaca, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1347–1356. URL: https://aclanthology.org/2024.findings-eacl.90/.

[17] H. Zhang, K. Ogasawara, Grad-CAM-based explainable artificial intelligence related to medical text processing, Bioengineering 10 (2023). URL: https://www.mdpi.com/2306-5354/10/9/1070. doi:10.3390/bioengineering10091070.

[18] W. Jin, X. Li, G. Hamarneh, Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 11945–11953. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21452. doi:10.1609/aaai.v36i11.21452.

[19] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, B. Zoph, GPT-4 technical report, 2023. URL: https://openai.com/research/gpt-4, OpenAI Technical Report.

[20] Y. Liu, Y. Li, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, L. Zhou, A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis, Meta-Radiology 2 (2024) 100099. URL: https://www.sciencedirect.com/science/article/pii/S2950162824000535. doi:https://doi.org/10.1016/j.metrad.2024.100099.

[21] P. Xu, X. Chen, Z. Zhao, D. Shi, Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis, British Journal of Ophthalmology 108 (2024) 1384–1389. URL: https://bjo.bmj.com/content/108/10/1384. doi:10.1136/bjo-2023-325054. arXiv:https://bjo.bmj.com/content/108/10/1384.full.pdf.

[22] S. Raza, D. J. Reji, F. Shajan, S. R. Bashir, Large-scale application of named entity recognition to biomedicine and epidemiology, PLOS Digital Health 1 (2022) e0000152. URL: https://doi.org/10.1371/journal.pdig.0000152. doi:10.1371/journal.pdig.0000152.

[23] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, T. Tong, HTC-Net: A hybrid CNN-transformer framework for medical image segmentation, Biomedical Signal Processing and Control 88 (2024) 105605. URL: https://www.sciencedirect.com/science/article/pii/S1746809423010388. doi:https://doi.org/10.1016/j.bspc.2023.105605.

[24] B. Vandersmissen, J. Oramas, On the coherency of quantitative evaluation of visual explanations, Computer Vision and Image Understanding 241 (2024) 103934. URL: https://www.sciencedirect.com/science/article/pii/S1077314224000158. doi:https://doi.org/10.1016/j.cviu.2024.103934.

[25] M. G. Ragab, S. J. Abdulkadir, A. Muneer, A. Alqushaibi, E. H. Sumiea, R. Qureshi, S. M. Al-Selwi, H. Alhussian, A comprehensive systematic review of YOLO for medical object detection (2018 to 2023), IEEE Access 12 (2024) 57815–57836. doi:10.1109/ACCESS.2024.3386826.

[26] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 7464 – 7475. doi:10.1109/cvpr52729.2023.00721.

[27] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of

Computer Vision 60 (2004) 91–110. doi:10.1023/B3AVISI.0000029664.99615.94.
[28] E. Rosten, R. Porter, T. Drummond, FASTER and better: A machine learning approach to corner detection, IEEE transactions on pattern analysis and machine intelligence 32 (2010) 105–19. doi:10.1109/TPAMI.2008.275.
[29] A. Budhkar, Q. Song, J. Su, X. Zhang, Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics, Computational and Structural Biotechnology Journal 27 (2025) 346–359. URL: https://www.sciencedirect.com/science/article/pii/S2001037024004495. doi:https://doi.org/10.1016/j.csbj.2024.12.027.
[30] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, A. Holzinger, Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists, Cognitive Systems Research 86 (2024) 101243. URL: https://www.sciencedirect.com/science/article/pii/S1389041724000378. doi:https://doi.org/10.1016/j.cogsys.2024.101243.
[31] J. Hatherley, L. Munch, J. Bjerring, In defence of post-hoc explanations in medical AI (2025). doi:10.48550.
[32] S.-C. Huang, M. Jensen, S. Yeung-Levy, M. P. Lungren, H. Poon, A. S. Chaudhari, Multimodal foundation models for medical imaging - a systematic review and implementation guidelines (2024). URL: https://www.medrxiv.org/content/early/2024/10/23/2024.10.23.24316003. doi:10.1101/2024.10.23.24316003.
[33] B. Antonio, D. Moroni, M. Martinelli, Efficient adaptive ensembling for image classification, Expert Systems 42 (2025) e13424. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13424. doi:https://doi.org/10.1111/exsy.13424. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13424.
[34] R. Zeineldin, M. Karar, Z. Elshaer, et al., Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI, Scientific Reports 14 (2024) 3713. URL: https://doi.org/10.1038/s41598-024-54186-7. doi:10.1038/s41598-024-54186-7.

Appendix

A. Online Resources

The result of this work is available via

- GitHub Repo: https://github.com/Jangulo7/med_explain_ja
- LLaVA-LLaMA 3 8B Fine-Tuned Model: https://huggingface.co/JoVal26/ja-med-clef-model
- LLaVA-Mistral 7B Fine-Tuned Model: https://huggingface.co/JoVal26/ja-clefmed-model

B. GRAD-CAM Image

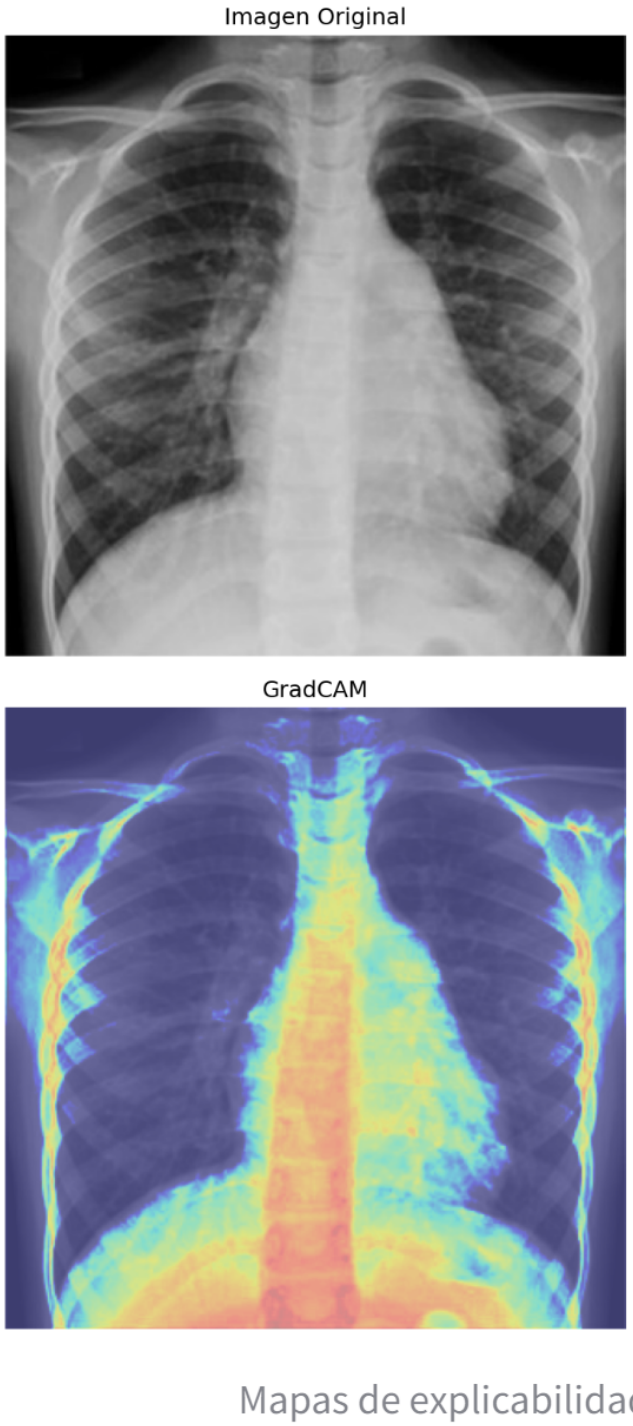

Figure 6: Sample GRAD-CAM visualization for image 32 of the test set, generated using an updated post-challenge version of the system. Source: Created by the authors using `ImageCLEFmedical_Caption_2025_test_32`, CC BY, Silva et al., 2024.