Modality-Guided Radiology Caption Prediction with Small Vision-Language Models and Image Classifier

Notebook for the CS_Morgan Lab at CLEF 2025

Raisa Nusrat Chowdhury[1], Mahmudul Hoque[1], Md Rakibul Hasan[1], Ejiga Peter Ojonugwa Oluwafemi[1] and Md Mahmudur Rahman[1,*]

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

Abstract

This working note describes our contributions to the ImageCLEFmedical 2025 Caption Prediction subtask, in which we investigated various approaches for extracting clinically relevant captions from radiological data. We made six unique submissions. Our study focused on modifying three vision-language models—Qwen-2B, Qwen2.5-3B, and SmolVLM-500M—using the ROCOv2 dataset, which contains radiological image-caption pairings. Three of our submissions employed direct caption generation, whereas the remaining three incorporated an additional image modality classification phase with a ResNet-50 model. The classifier output (e.g., CT, MRI, Ultrasound, Radiograph) was included into the prompt to enhance caption generating efficacy. Among all submissions, Qwen 2B (2B params) emerges as the strongest overall performer, achieving the best scores in Overall (0.2316), Similarity (0.5704), ROUGE (0.1598), Relevance (0.3717), UMLS Concepts F1 (0.0741), AlignScore (0.1087), and Factuality Average (0.0914). These results indicate that Qwen 2B is highly effective at producing clinically accurate and factually aligned captions. In contrast, SmolVLM (Classification) leads in BERTScore (0.5375) and BLEURT (0.2576), suggesting that it excels in capturing semantic meaning and producing fluent, human-like text. These complementary strengths reflect different modeling priorities: Qwen 2B focuses more on factual and structural alignment, while SmolVLM emphasizes linguistic similarity and coherence. These findings underscore the effectiveness of hybrid pipelines that combine classification with prompt adaptation. Moreover, they show that even resource-efficient models, when fine-tuned and guided properly, can provide clinically valuable outputs. Our experiments support the ongoing shift toward smaller, adaptable vision–language systems for medical AI, offering practical potential for deployment in low-resource settings.

Keywords

Vision Language Model, Medical Image Captioning, Qwen2.5-VL, Qwen2-VL, SmolVLM

1. Introduction

Radiological imaging has become much more common in recent years, which has made clinical processes very difficult. Due to more cross-sectional studies (i.e., CT and MRI) and more complex images, diagnostic radiologists' overall workload "has increased considerably," according to studies [1]. A review found that 97% of UK imaging units couldn't keep up with clinical demand. This means that many hospitals are always short of staff [2]. Writing up thorough radiology reports by hand takes a lot of time and could cause delays or changes in how patients are cared for. These problems have made people want to automate some parts of the reporting process. In this case, image captioning that is driven by AI has a lot of potential [3]. As a vision–language task, automated radiology report generation (ARRG) is similar to image captioning. It has been shown to "have significant clinical value and could alleviate time pressures" by taking care of routine [4]. Modern vision–language foundation models (VLMs), pre-trained on large-scale image–text datasets, have demonstrated remarkable capabilities [5]. These

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

*Corresponding author.

racho1@morgan.edu (R. N. Chowdhury); mahoq1@morgan.edu (M. Hoque); mdhas1@morgan.edu (M. R. Hasan); ejiga.ojonugwa.peter@gmail.com (E. P. O. Oluwafemi); md.rahman@morgan.edu (M. M. Rahman)

https://github.com/raisanusrat222 (R. N. Chowdhury); https://github.com/HoqueMahmudul (M. Hoque); https://github.com/Hasan-MdRakibul (M. R. Hasan)

0000-0002-1663-6391 (R. N. Chowdhury); 0009-0006-5532-4135 (M. Hoque); 0000-0002-6179-2238 (M. R. Hasan); 0009-0003-2039-3075 (E. P. O. Oluwafemi)

models effectively learn complex associations between visual and textual information and are able to perform zero-shot recognition and description across a broad range of tasks. The fact that these models did well in general-domain vision–language tests suggests that they could be used in medical imaging after being fine-tuned on data specific to that field [6]. This study investigates the application of big Vision-Language Models for medical image captioning. The CS_Morgan team engaged in the ImageCLEFmedical 2025 Caption Prediction subtask, concentrating on the generation of clinically relevant captions from radiological images. We optimized multiple cutting-edge VLMs such as Qwen-2B [7], Qwen2.5-3B [8], and SmolVLM-500M [9] using the supplied training data. Certain iterations of our methodology also included auxiliary modules, such as an image-modality classifier, and employed prompt-tuning procedures to enhance relevance.

2. Objectives

As part of the ImageCLEFmedical 2025 Caption Prediction task [10, 11], this work investigates and assesses the potential of vision-language models to produce clinically relevant captions for medical images. The main goals are:

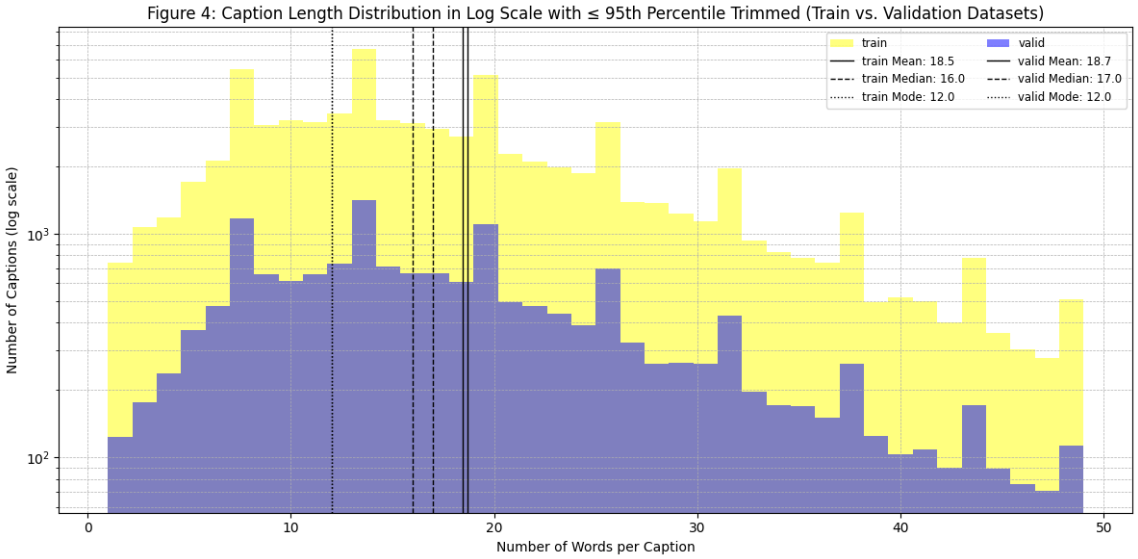
1. To use advanced vision-language models to produce medically appropriate and descriptive captions for radiology images from various modalities, including MRI, CT, X-ray, and Ultrasound.
2. To assess the effectiveness of various VLM's configurations using performance metrics (e.g., BERTScore, ROUGE, BLEURT, Relevance, UMLS Concepts F1, AlignScore, and Factuality Average).
3. To examine, using a unified evaluation framework, the generalization potential of different model architectures, including large-scale language-vision models and lightweight alternatives.
4. To provide the research community with comparative insights into model selection and fine-tuning strategies for medical image captioning.

3. Dataset

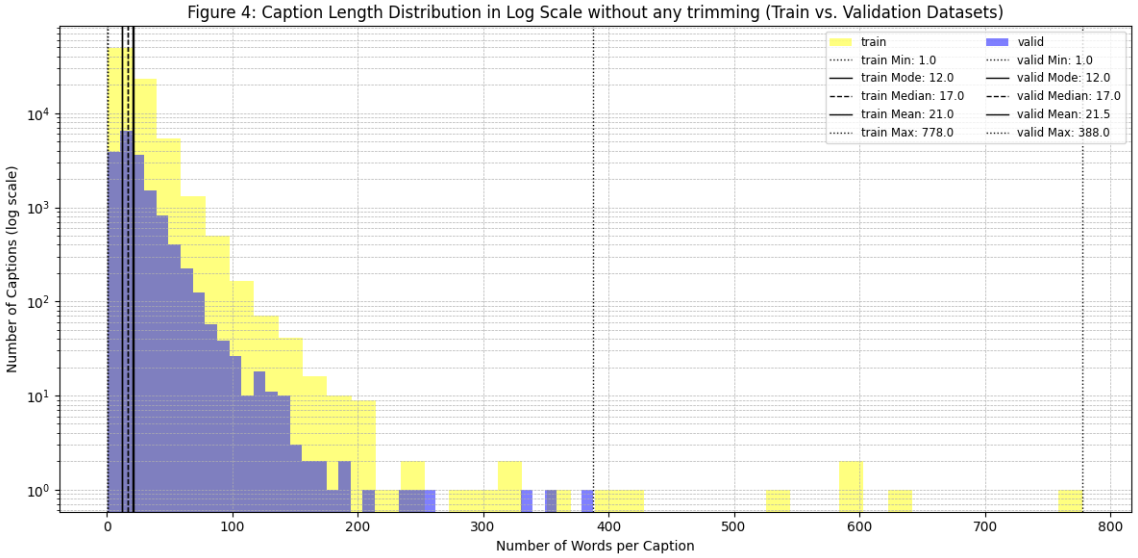
All of the models mentioned in the introduction section were trained and assessed using the ROCOv2 [12] dataset supplied by the competition organizers [10, 11]. The dataset comprises about 80,091 training, 17,277 validation, and 19,267 test radiology images accompanied by captions and concepts. This dataset consists of many modalities and anatomical locations, providing an extensive platform for biomedical captioning. Our objective is to evaluate the efficacy of contemporary VLMs in producing precise, therapeutically relevant captions within the biomedical field. Figure 1 provides a comprehensive view of the dataset's linguistic characteristics. **Subfigure (a)** shows that the majority of captions are between 10 and 30 words long, with aligned mean, median, and mode across train and validation splits, indicating a stable annotation style. **Subfigure (b)** highlights the long-tailed nature of the dataset, revealing rare captions with more than 500 words. Despite these extremes, central tendencies remain consistent. **Subfigure (c)** displays the most frequent words used in the captions after preprocessing (i.e., removing punctuations, stop words, numeric words, etc.) Together, these subfigures support the inference that the dataset is well-structured, domain-specific, and linguistically consistent — ideal for training and evaluating medical image captioning models.

4. Models to Predict Captions

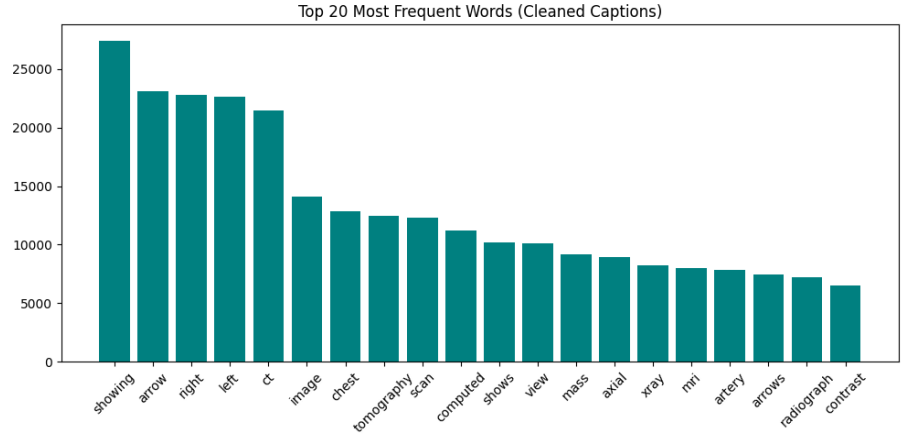
The models employed here are trained to predict captions for the radiology images. This process consists of three steps: (1) utilizing ResNet50 to categorize the images into four major image modalities such as CT, MRI, Ultrasound, and Radiograph, (2) training of SmolVLM, Qwen-2B, and Qwen-2.5-3B with some parameter efficient techniques to predict captions, and (3) using structured prompt engineering to guide the trained models regarding the predicted modality of test images to generate more relevant



(a) Caption length distribution (log scale) with $\leq$ 95th percentile trimmed.



(b) Caption length distribution (log scale) without trimming.



(c) Top 20 most frequent words in cleaned captions.

Figure 1: Statistical analysis of medical image captions: (a) shows the distribution of caption lengths with 95th percentile trimming; (b) includes the full length range with outliers; and (c) presents the most frequent domain-relevant words across all captions.

captions. Figure 2 depicts the corresponding framework and the following sections will elaborate the relevant parts of this framework.

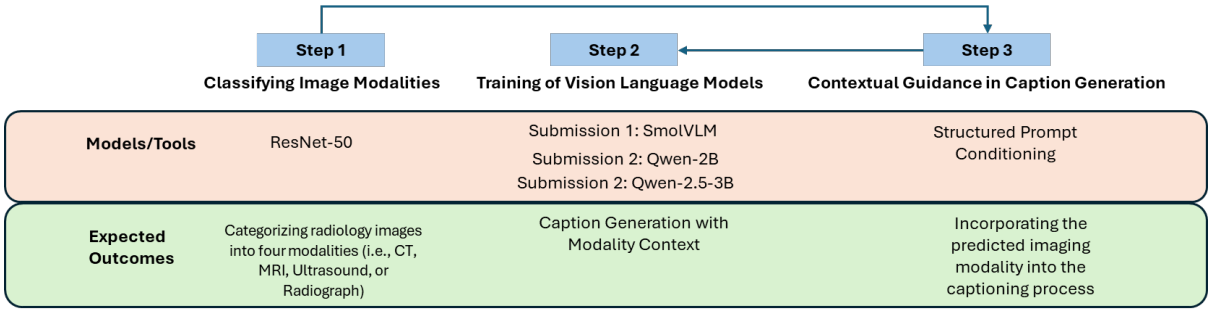


Figure 2: Conceptual Framework of the Experiments

4.1. SmolVLM

The language model is a simple SmolLM decoder with 500 million parameters. Advanced efficiency features like Grouped-Query Attention and FlashAttention allow it to handle up to 4,096 token sequences and provide fast inferences by default. A two-layer MLP projection matrix smooths the integration of visual and textual input from 1,024-dimensional vision tokens into the language model's 4,096-dimensional embedding space. The language model, a streamlined SmolLM decoder with 500 million parameters, inherits additional efficiency features like Grouped-Query Attention and FlashAttention to support large sequences (up to 4,096 tokens by default) and quick inference [9]. A two-layer MLP projection matrix transforms the 1,024-dimensional vision tokens into the 4,096-dimensional embedding space employed by the language model. This ensures that visual and written information can be utilized concurrently without any issues. The lightweight SigLIP encoder, in conjunction with the projection layer and the distilled SmolLM decoder, enables SmolVLM-500M to perform competitively in captioning and visual question answering, requiring approximately 1.2 GB of GPU memory for a complete image-to-text conversion. This renders it ideal for deployment on consumer-grade devices.

4.2. Qwen2-VL and Qwen2.5-VL

Qwen2-VL [13] uses a dynamic-resolution ViT encoder to process images of any size into visual embeddings, which are then mapped into a 4096-dimensional language space using a two-layer MLP. Its decoder is a 32-layer Qwen-2B model with 2B parameters, employing Grouped-Query Attention and Flash-style kernels to handle sequences up to 8,192 tokens efficiently. This architecture enables fine-grained image reasoning and coherent multimodal generation, all deployable on a single GPU. **Qwen2.5-VL** [14] builds upon this by using a ViT encoder trained with Naïve Dynamic Resolution and enhanced with 2D positional encodings. The same MLP structure projects vision tokens to match the 4096-dimension text space. It incorporates a more powerful 36-layer Qwen-2.5 decoder with over 3B parameters, capable of handling up to 32,768 tokens. Despite its increased capability—supporting detailed grounding and complex outputs—it remains efficient enough for single-GPU deployment.

5. Submissions for the Caption Prediction Task

5.1. Submission 1: LoRA fine-tuning of SmolVLM-500M

This submission modifies the publicly available SmolVLM-500M-Instruct checkpoint—a 12-layer `ViT-B` vision encoder paired with a 32-layer text decoder—to produce captions for ROCO-v2 radiological images. Rather than updating all 500 million parameters, we utilized LoRA on each linear transformation within both the vision and text stacks. The language model's projection head (`lm_head`) and token

embeddings (embed_tokens) were also kept trainable. The technique combines low-rank adaptation with selectively updated weights to capture domain-specific vocabulary while reducing memory usage and overfitting risks.

LoRA was configured with a rank of $r = 64$, a scaling factor of $\alpha = 16$, and a dropout rate of 0.10 to regularize the adapter pathways. The bias terms were fixed (bias="none"), while the use_rslora flag enabled the more memory-efficient RS-LoRA formulation. The adapters added around 7.4 million trainable parameters, constituting about 1.5% of the model's overall size, enabling the entire process to be efficiently executed on a single A100-80 GB GPU with bfloat16 precision.

5.1.1. Training Process

Training was executed with the Transformers Trainer, gradient check-pointing, and FlashAttention to optimize memory use. The modality-balanced ROCO-v2 split underwent processing for two complete epochs (about 5,006 optimization steps) with an effective batch size of 32 (per-device batch of 2, gradient accumulation of 16). Optimization utilized AdamW ($\beta = 0.9/0.999$, $\epsilon = 1 \times 10^{-8}$) with a base learning rate of 5×10^{-5}, following a cosine decay schedule and included 500 warm-up steps. Checkpoints and evaluations were documented after each epoch; the first-epoch checkpoint—training loss = 2.5229, validation loss $\approx$ 2.58597—exhibited the minimal loss values and was hence selected for downstream inference. The adapters from this checkpoint were integrated with the fixed backbone, and captions were produced using greedy decoding (temperature = 1.0, num_beams = 1, max_new_tokens = 64).

5.2. Selective fine-tuning of Qwen-2B

For this run, the Qwen-2 Vision-Language checkpoint with two billion parameters was loaded. The backbone connects a 32-layer vision transformer (hidden = 1,280) to a 28-layer text decoder (hidden = 1,536) using a trained multimodal projector. LoRA was used to achieve parameter-efficient adaptation, with a rank $r = 32$, lora_alpha = 32, and dropout rate = 0.10. Adapters were injected into each self-attention block's key (k_proj), query (q_proj), value (v_proj), and output (o_proj) projections, as well as the up, down, and gate projections (up_proj, down_proj, gate_proj) of each MLP. The identical set of projections in the vision attention blocks, as well as both fully connected layers of the vision MLP (mlp.0 and mlp.2), were addressed. Furthermore, the language head (lm_head) and input embeddings (embed_tokens) were left trainable to provide domain-specific vocabulary adaption. This selective technique activated 504,217,600 parameters out of 2,713,203,200, which means that approximately 18.6% of the model can be updated during training.

5.2.1. Training Process

To reduce memory usage, training was undertaken using the Transformers Trainer, gradient check-pointing (use_reentrant = False), and Flash-Attention. The modality-balanced ROCO-v2 split ran through the data three times, recording and saving every 500 optimization steps; batches were streamed in bfloat16 with a per-device size of one, and no gradient accumulation. Optimization used AdamW ($\beta = 0.9/0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay = 0.01) with a base learning rate of 3×10^{-4} and a typical cosine schedule. The validation loss fell from 4.41 at step 500 to a low of 4.18 at step 2,500 (with a corresponding training loss of around 3.74); this checkpoint was automatically recognized as the best. The adapters from checkpoint-2500 were combined with the frozen backbone for inference, and captions were created using greedy decoding (temperature = 1.0, num_beams = 1, max_new_tokens = 64).

5.3. Selective fine-tuning of Qwen-2.5-3B

This experiment utilized the multimodal backbone Qwen-2.5-VL-3B, which integrates a 32-layer vision transformer with a 36-layer causal text decoder including 2,048 hidden units. Parameter-efficient adap-

tation was achieved by Low-Rank Adaptation. Rank-32 adapters, configured with a lora_alpha of 16 and regularized by a 0.10 dropout, were incorporated into every query, key, value, and output projection of the self-attention blocks, as well as the up_, down_, and gate_projection pathways of each MLP in both the vision and text stacks. The identical adapter template was utilized for the two fully connected layers within the vision-side MLP. Furthermore, the language-model head (lm_head) and the token-embedding matrix (embed_tokens) were retained as trainable components to enable the model to enhance its medical lexicon. This selective technique optimized 702,435,328 weights—approximately 15.76% of the 4,457,058,304 parameters in the original checkpoint—while the remaining layers remained fixed in bfloat16.

5.3.1. Training Process

Training was conducted via the Transformers Trainer, with gradient check-pointing (re-entrant path disabled) and Flash-Attention kernels. Caption pairings from the modality-balanced ROCO-v2 corpus were transmitted to the GPU in bfloat16 format for four notional epochs; each optimization step handled a single image-caption pair, as both the training and evaluation batch sizes were configured to one, and no gradient accumulation was employed. AdamW was used for optimization using a base learning rate of 3×10^{-4}, cosine decay scheduling, and a weight decay coefficient of 0.01. Every 500 optimization steps, a full evaluation and check-pointing were done. An early-stopping callback monitored the validation loss with a threshold of 0.01 and a patience of three evaluations.

The first evaluation at step 500 showed a training loss of 4.52 and a validation loss of 4.09. Later evaluations indicated a continuous decline in validation loss, reaching its lowest point of 3.79 at step 5,000, at which the best-model flag was set. As a result, the adapters saved in checkpoint-5000 were used for subsequent caption generation. This was done using greedy decoding with a temperature of 1.0, a single beam, and a maximum of 64 new tokens.

5.4. Captioning with Classification

In the ROCOv2 dataset, the four major imaging modalities—CT (28,005), MRI (12,669), Radiograph (26,789), and Ultrasound (11,425)—were chosen because they are the categories that occur the most frequently and have the most significant diagnostic implications. There were 1,203 images that were labeled as "Other," This included some other modalities, such as nuclear scans and positron emission tomography (PET). They could have caused class imbalance and instability in the modeling activities that followed.

To prevent class imbalance, each modality's training set was downsampled to match the smallest class size, ensuring equal representation for robust learning on the imbalanced dataset [15]. Cross-entropy loss and validation oversight were used throughout the ResNet classifier's entire training phase, which lasted five epochs from the start. When each class was downsampled, it was assigned to the category that was the smallest. This ensured that all classes were represented equally. Although the training period was relatively short, the classifier was able to achieve stable validation performance, with accuracy, precision, and recall reaching 96.46%. This indicates that the model was successful in capturing modality-specific properties. In light of the fact that test labels were not available at the time of submission, the evaluation was limited to validation metrics. The classifier's predictions for test images were subjected to distributional sanity checks, which ensured that the output was balanced across all modalities.

Figure 3 shows the overview of the proposed hybrid pipeline integrating classification and prompt-based captioning.

A ResNet-50 convolutional neural network [16], trained from scratch without ImageNet pretraining, was adapted to output four modality classes. The model was trained for five epochs using cross-entropy loss and optimizers such as Adam or SGD, with checkpoints saved each epoch and validation monitoring to avoid overfitting [17].

The classifier achieved high validation accuracy (96.46%) and balanced precision, recall, and F1-scores

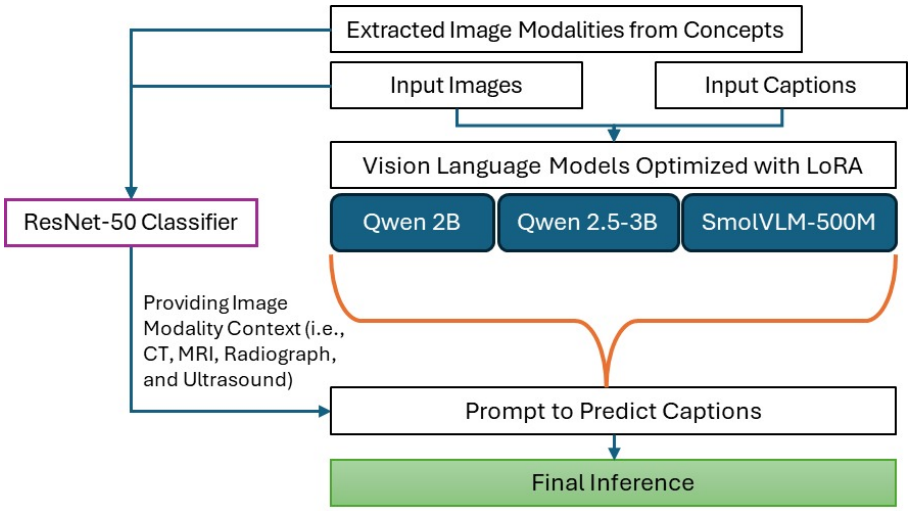


Figure 3: Framework for Medical Image Captioning Using Modality-Informed Vision-Language Models

(approximately 96.4%), indicating effective feature learning and minimal bias. Finally, the trained model predicted modalities on the test set, outputting results as a CSV file for subsequent caption generation.

5.5. Contextual Guidance in Caption Generation

Structured Prompt Conditioning: The ResNet-50 classifier generated predicted modality labels, which were then saved to a CSV file containing image identifiers and modality categories. During inference, each captioning model determined the predicted modality for a given test image and included it in the input prompt. For example, prompts were formatted as follows: "CT image: Describe the medical image." This modality-aware conditioning helped the vision-language models generate more context-specific and clinically acceptable captions. Importantly, modality information was employed only during inference, not model training. This method enabled all three models—SmolVLM-500M, Qwen-2B, and Qwen2.5-3B—to remain modular and flexible while retaining modality-specific vocabulary and structure in their outputs.

Caption Generation with Modality Context: During inference, each captioning model processes the input image along with the modality-conditioned prompt to generate a relevant textual description. Conditioning on the modality enables the use of modality-specific vocabulary—such as ultrasound terms for ultrasound images or scanning sequence references for MRI—reducing ambiguity and ensuring modality-accurate captions.

All three captioning models (Qwen-2B, Qwen2.5-3B, and SmolVLM-500M) employ this structured guidance, analyzing both the modality-conditioned prompt and image simultaneously. This two-stage pipeline—modality classification followed by modality-aware captioning—effectively adapts general vision-language models to produce clinically relevant, modality-consistent medical image captions.

6. Evaluation Metrics

Overall Average: This cumulative score is the mean of all individual metric scores for a submission, providing a comprehensive assessment of caption quality in terms of relevance and factual accuracy. A higher Overall Average indicates better performance and was used by organizers to rank systems.

Image-Caption Similarity: Measures semantic consistency between image content and caption by embedding both via a medical image-text model and computing their similarity. A high score reflects accurate depiction beyond simple lexical overlap.

BERTScore [18] evaluates text similarity by comparing BERT embeddings of candidate and reference captions, capturing semantic meaning more effectively than traditional methods.

ROUGE [19] assesses overlap of n-grams, word sequences, and pairs between candidate and reference texts, commonly used for summarization and translation evaluation.

BLEURT [20] is a learned metric leveraging fine-tuned pretrained transformers to predict human judgment scores for natural language generation.

Relevance Average: The average of Image-Caption Similarity, BERTScore, ROUGE-1, and BLEURT, measuring how well captions match reference content and language. A high score indicates informative and on-topic captions.

UMLS Concept F1: Evaluates factual overlap of clinical concepts by extracting medical entities from generated and reference captions using UMLS [21]. The F1-score compares true positives, false positives, and false negatives of these entities.

AlignScore: Uses a RoBERTa-based verifier to break captions into claims and verify their support by the image context [22]. This checks true factual alignment rather than mere word overlap, with higher scores indicating better claim support.

Factuality Average: The mean of UMLS Concept F1 and AlignScore, reflecting overall factual accuracy by balancing term-level concept overlap and sentence-level claim verification.

7. Results and Discussions

In order to explore the effects of model capacity, modality conditioning, and inference design on semantic and clinical performance, our team assessed six captioning techniques. The first two entries used vision-language backbones that had already been trained on medical image-caption pairings, Qwen 2B and Qwen 2.5-3B. Qwen 2B (Submission 3) obtained the best overall composite score (0.2316), despite having a higher BERTScore (0.5296), which is probably because to its larger decoder size. Inadequate factual basis was demonstrated by the poor Factuality Averages of both models (e.g., 0.0601 for Qwen 2.5–3B). Submissions 4 and 5 presented a two-stage methodology wherein each test image was initially categorized into one of four primary imaging modalities—CT, MRI, Radiograph, or Ultrasound—utilizing a ResNet-50 classifier. The anticipated label was subsequently employed to build a structured input prompt, for instance, "MRI image: Describe the medical image." This directed the captioning model towards language relevant to the modality. The independently trained classifier, utilized solely during inference, had a notable impact: Qwen 2.5 with classification (Submission 4) enhanced BLEURT from 0.2040 to 0.2263 and UMLS Concepts F1 from 0.0225 to 0.0271, in contrast to its non-classification equivalent. These advancements indicate both semantic enhancement (assessed by BLEURT, which evaluates linguistic fluency and contextual alignment) and clinical significance (measured by UMLS Concept F1, which quantifies the intersection of medically relevant ideas).

Qwen 2B with classification (Submission 5) shown robust performance across all metrics, enhancing both Relevance Average (from 0.2853 to 0.3132) and Factuality Average (from 0.0351 to 0.0523). These enhancements underscore the advantages of modality-aware prompting in augmenting both the interpretability and clinical accuracy of generated captions.

Submissions 1 and 6 employed the smaller SmolVLM-500M model to investigate performance-efficiency trade-offs. The baseline model (Submission 1) attained substantial semantic alignment (BERTScore 0.5361, BLEURT 0.2518), although exhibited diminished factual content (Factuality 0.0380). The implementation of classification-based prompts in Submission 6 enhanced BLEURT to 0.2576 and achieved the highest Relevance Average among tiny models at 0.3646, demonstrating that even lightweight architectures can gain from modality conditioning.

Significantly, these classification-driven improvements were attained without altering the fundamental VLM settings. By implementing predicted modality labels solely during inference, we maintained the system's flexibility and modularity. This design decision was intentional: including modality labels during training would necessitate architectural alterations and re-tuning of model weights, hence augmenting complexity. Utilizing classifier outputs as prompts facilitates controlled, domain-specific captioning while maintaining efficiency—an advantageous characteristic for practical clinical use when minimal updates are desired.

Although we did not perform formal ablation tests, the comparative results of each model with and without classifier assistance offer implicit proof of the technique's influence. Modality-informed prompting enhanced semantic overlap (BLEURT, BERTScore), clinical correctness (UMLS F1), or structured contextual alignment (Relevance Average) across all evaluated backbones. These findings indicate that hybrid pipelines integrating independent classifiers with optimized VLMs provide a scalable and efficient approach for medical caption production. Future research will examine the comparative efficacy of training-time integration of modality features against our inference-only configuration, as well as the potential benefits of combined optimization methodologies.

Table 1
Evaluation metrics for captioning submissions

Sub #	Model	Overall	Similarity	BERTScore	ROUGE	BLEURT	Relevance	UMLS Concepts F1	AlignScore	Factuality Average
3	Qwen 2B (2B params)	**0.2316**	**0.5704**	0.5180	**0.1598**	0.2385	**0.3717**	**0.0741**	**0.1087**	**0.0914**
2	Qwen 2.5-3B	0.1882	0.4456	0.5296	0.0873	0.2026	0.3163	0.0244	0.0956	0.0601
6	SmolVLM (Classification)	0.1928	0.4202	**0.5375**	0.1361	**0.2576**	0.3646	0.0233	0.0725	0.0479
4	Qwen 2.5 (Classification)	0.1819	0.4507	0.5111	0.0919	0.2040	0.3512	0.0225	0.0762	0.0494
5	Qwen 2B (Classification)	0.1827	0.3925	0.5089	0.1347	0.2164	0.3132	0.0171	0.0874	0.0523
1	SmolVLM	0.2001	0.4143	0.5361	0.1362	0.2518	0.3622	0.0157	0.0602	0.0380

8. Conclusion

This study explored a modular approach to medical image captioning by fine-tuning vision-language models and incorporating image modality prompts during inference. The models included Qwen-2B, Qwen-2.5-3B, and SmolVLM-500M, each fine-tuned on radiology captioning tasks. A ResNet-50 classifier was trained to predict the modality of each input image, and these modality labels were included as prompts during caption generation to provide contextual guidance. The introduction of modality prompts did not improve overall captioning performance across all models. However, it led to measurable gains in factual accuracy and medical concept coverage for smaller model. For example, the SmolVLM baseline achieved a Factuality Average of 0.0380 and UMLS Concept F1 of 0.0157, while the modality-informed variant reached 0.0479 and 0.0223. These results suggest that even minimal contextual signals can help smaller models better align with domain-specific terminology and clinical details. The framework prioritized simplicity and modularity by limiting the classifier to inference and focusing on major modality categories. This structure allowed for efficient reuse of components and rapid experimentation without requiring retraining of the core models. Future directions include incorporating modality information during training, expanding the modality taxonomy, and conducting component-level ablation studies. The findings suggest that contextual guidance provides a practical approach for enhancing the specificity and factual accuracy of generated captions in clinical applications.

9. Acknowledgments

This work was supported by the National Science Foundation (NSF) grant (ID. 2131307) "CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support."

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using the mentioned tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. C. Kwee, R. M. Kwee, Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence, Insights into imaging 12 (2021) 1–12.

[2] A. Rimmer, Radiologist shortage leaves patient care at risk, warns royal college, BMJ: British Medical Journal (Online) 359 (2017).

[3] O. O. E. Peter, M. M. Rahman, F. Khalifa, Advancing ai-powered medical image synthesis: Insights from medvqa-gi challenge using clip, fine-tuned stable diffusion, and dream-booth+ lora, arXiv preprint arXiv:2502.20667 (2025).

[4] P. Sloan, P. Clatworthy, E. Simpson, M. Mirmehdi, Automated radiology report generation: A review of recent advances, IEEE Reviews in Biomedical Engineering (2024).

[5] O. O. E. Peter, A. Oluwapemiisin, A. Chetachi, A. Opeyemi, F. Khalifa, M. M. Rahman, Synthetic data-driven multi-architecture framework for automated polyp segmentation through integrated detection and mask generation, in: Medical Imaging 2025: Clinical and Biomedical Imaging, volume 13410, SPIE, 2025, pp. 558–569.

[6] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[7] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023. URL: https://arxiv.org/abs/2308.12966. arXiv:2308.12966.

[8] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, Z. Fan, Qwen2 technical report, 2024. URL: https://arxiv.org/abs/2407.10671. arXiv:2407.10671.

[9] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. B. Allal, A. Lozhkov, N. Tazi, V. Srivastav, J. Lochner, H. Larcher, M. Morlon, L. Tunstall, L. von Werra, T. Wolf, SmolVLM: Redefining small and efficient multimodal models, 2025. URL: https://arxiv.org/abs/2504.05299. arXiv:2504.05299.

[10] 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 ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[11] B. Ionescu, H. Müller, D. Stanciu, A. Andrei, A. Radzhabov, Y. Prokopchuk, L. Ştefan, M. 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. 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.

[12] 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.

[13] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang,

M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, 2024. URL: https://arxiv.org/abs/2409.12191. arXiv:2409.12191.

[14] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL Technical Report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[15] M. Hassan, S. Ali, H. Alquhayz, K. Safdar, Developing Intelligent Medical Image Modality Classification System using Deep Transfer Learning and LDA, Scientific Reports 10 (2020) 12868.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Computer Vision and Pattern Recognition, 2015, pp. 770–778. doi:10.1109/cvpr.2016.90.

[17] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?, IEEE Transactions on Medical Imaging 35 (2016) 1299–1312. URL: http://dx.doi.org/10.1109/TMI.2016.2535302. doi:10.1109/tmi.2016.2535302.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, International Conference on Learning Representations abs/1904.09675 (2019).

[19] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[20] T. Sellam, D. Das, A. P. Parikh, BLEURT: Learning Robust Metrics for Text Generation, in: Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892. doi:10.18653/v1/2020.acl-main.704.

[21] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. Folarin, A. Roberts, R. Bendayan, M. Richardson, R. Stewart, A. Shah, W. K. Wong, Z. M. Ibrahim, J. Teo, R. Dobson, Multi-Domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit, Artificial Intelligence in Medicine 117 (2020) 102083. doi:10.1016/j.artmed.2021.102083.

[22] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, D. Jurafsky, Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation, in: NAACL-HLT, 2021, pp. 5288–5304. doi:10.18653/v1/2021.naacl-main.416.