

UIT_HighDefinition at ImageCLEFmedical Caption 2026: Soft CUI-Guided Qwen3-VL Captioning

Notebook for the DS-MED Lab (UIT) at CLEF 2026

Nguyen Tan Duy^{1,2}, Nguyen Thi Quynh Han^{1,2}, Nguyen Minh Triet^{1,2} and Thien B. Nguyen-Tat^{1,2,*}

¹University of Information Technology, Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

Purpose: This paper presents the UIT_HighDefinition system submitted to the ImageCLEFmedical Caption 2026 Caption Prediction task. The task requires generating concise and clinically meaningful captions for heterogeneous biomedical images.

Method: The submitted approach treats the task as instruction-guided medical image caption generation with soft biomedical concept conditioning. Qwen3-VL-8B-Instruct is used as the caption generator, and predicted Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) are incorporated as optional clinical hints. These auxiliary CUI predictions were generated by running the public UIT_Worker Concept Detection notebook for ImageCLEFmedical 2026; the Concept Detection training and inference code was not modified in this work. The predicted CUIs are mapped to readable biomedical terms. Instead of forcing all predicted concepts into the output, the final prompt instructs the model to use CUI terms only when visually supported. A conservative rule-based cleaner normalizes modality expressions, removes generic prefixes, controls length, and validates submission formatting. Inference is distributed across six GPU workers with rank-level validation and resume support.

Results: The final official submission achieved an overall score of 0.3193 and ranked seventh on the Caption Prediction Test 2026 leaderboard. The official run was reported under the final team name UIT_HighDefinition with run ID 980; the original leaderboard username was tanduy2702. It obtained scores of 0.4999 for relevancy, 0.1388 for factuality, 0.5977 for BERTScore, 0.2243 for ROUGE, 0.8496 for semantic similarity, 0.3278 for BLEURT, 0.1397 for MedCAT, and 0.1379 for alignment. Compared with the Qwen2.5-VL-3B image-only baseline, the final system improved the overall score from 0.2934 to 0.3193.

Conclusion: The results suggest that large vision-language models can produce useful medical image captions when combined with soft concept-level guidance and careful output validation. However, the gap between high semantic similarity and lower MedCAT/alignment scores indicates that exact biomedical grounding remains challenging.

Keywords

Medical image captioning, ImageCLEFmedical Caption, Vision-language model, Qwen3-VL, UMLS, CUI-guided prompting, Biomedical concept grounding

1. Introduction

Medical image captioning aims to automatically generate textual descriptions that summarize the visual and clinical content of biomedical images. This task is important because biomedical images are widely used in scientific articles, clinical documentation, education, and diagnostic workflows. A reliable captioning system can support physicians and researchers by producing concise descriptions of imaging modalities, anatomical regions, abnormalities, and visible findings. Nevertheless, medical image captioning remains difficult because biomedical figures are heterogeneous and may include X-ray,

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

*Corresponding author.

✉ 24520389@gm.uit.edu.vn (N. T. Duy); 24520013@gm.uit.edu.vn (N. T. Q. Han); 24521852@gm.uit.edu.vn (N. M. Triet); thienntb@uit.edu.vn (T. B. Nguyen-Tat)

🆔 0009-0003-1969-3388 (N. T. Duy); 0009-0000-0042-8523 (N. T. Q. Han); 0009-0009-4476-9472 (N. M. Triet); 0000-0002-4809-7126 (T. B. Nguyen-Tat)



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

computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, endoscopy, histology, microscopy, angiography, hybrid imaging, and multi-panel scientific figures.

Compared with general-domain image captioning, medical caption generation requires stronger domain knowledge and more careful language control. A generic caption such as “a medical image” is not sufficient; the system is expected to identify the modality, anatomical region, and visible finding when possible. At the same time, hallucinating unsupported diseases or patient information is particularly problematic in the medical domain. These challenges motivate the use of large vision-language models (VLMs) combined with domain-specific guidance and conservative post-processing.

ImageCLEFmedical Caption is a long-running benchmark for medical image understanding. The ImageCLEF 2026 overview provides the broader evaluation context for the multimodal CLEF tasks [1]. The ImageCLEFmedical 2026 task overview defines the medical concept detection and caption generation setting used in this work [2]. Previous ImageCLEFmedical Caption editions evaluated concept detection and caption prediction on biomedical figures and captions [3, 4, 5].

Recent teams have explored several methodological directions. Encoder-decoder captioning and multi-label concept prediction were used in earlier ImageCLEFmedical Caption submissions [6, 7]. BLIP-based and Transformer-based systems were explored by UIT-2Q2T and UIT-DarkCow [8, 9]. Concept-aware captioning and attention-based concept integration were investigated by DS@BioMed [10]. Retrieval-augmented and reranking-based approaches were also studied in recent ImageCLEFmedical Caption systems [11, 12]. These studies show that medical image captioning benefits from strong multimodal backbones, domain-specific concepts, and careful validation.

This paper describes the UIT_HighDefinition team’s system for the ImageCLEFmedical Caption 2026 Caption Prediction task. The system uses Qwen3-VL-8B-Instruct as the main caption generator [13]. To inject biomedical knowledge, predicted Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) are converted into textual terms and included in the prompt as optional clinical hints. Unlike hard concept forcing, the prompt explicitly instructs the model to use these terms only when visually supported. This design aims to improve medical terminology coverage while reducing unsupported clinical claims.

The main contributions of the submitted system are:

- a Qwen3-VL-8B-based medical image captioning pipeline for ImageCLEFmedical Caption 2026;
- a visually grounded soft CUI-guidance strategy that uses predicted UMLS concepts as optional prompt hints rather than mandatory output labels;
- the use of externally generated CUI predictions as soft clinical prompt hints for Qwen3-VL captioning;
- a conservative rule-based cleaner for modality normalization, prefix removal, length control, and final submission validation;
- a fault-tolerant six-rank distributed inference workflow for processing all 15,249 test images.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the task and data. Section 4 presents the proposed system. Section 5 reports implementation details and results. Section 6 discusses observations and limitations. Section 7 concludes the paper.

2. Background and Related Work

2.1. Medical image captioning and concept detection

Medical image captioning combines computer vision and natural language generation. The ImageCLEFmedical Caption benchmark has become one of the main evaluation settings for this problem. The 2026 task overview defines the current medical concept detection and caption generation setting [2], while previous task overviews describe earlier Caption Prediction and Concept Detection editions [3, 4, 5]. ROCov2, an updated and extended version of the Radiology Objects in Context dataset, has been widely used in recent editions of the task [14].

Early competitive systems often used encoder-decoder architectures, where visual features are extracted by an image encoder and captions are generated by a language decoder. CSIRO proposed a compact medical image captioning model for ImageCLEFmedical Caption 2023 [6], while the VCMi team explored concept detection and caption prediction using multi-label classification and generation methods [7]. In the 2024 edition, DBS-HHU studied both concept detection and caption prediction, using convolutional neural network (CNN) ensembles for concept detection and the Generative Image-to-Text Transformer (GIT) for caption generation [15]. UIT-DarkCow investigated Transformer encoder-decoder and Query Transformer architectures [9], UIT-2Q2T used Bootstrapping Language-Image Pre-training (BLIP) for medical image captioning [8], and DS@BioMed integrated concept detection into caption generation through attention-based mechanisms [10].

Concept detection is closely related to captioning because medical captions often require precise biomedical terms. UMLS provides a unified biomedical terminology resource that links many vocabularies and concept identifiers [16]. Prior work indicates that concept information can improve the specificity of generated captions, but noisy concepts may also mislead the generator. The proposed approach therefore uses CUIs as soft prompt-level hints rather than hard constraints.

2.2. Vision-language models and prompt-based medical captioning

Vision-language pretraining has substantially improved image captioning. Bootstrapping Language-Image Pre-training (BLIP) [17], BLIP-2 [18], and InstructBLIP [19] introduced strong vision-language pretraining and instruction-following capabilities. In ImageCLEFmedical Caption 2025, several systems used InstructBLIP or related VLMs as captioning backbones [20, 11, 12]. DS4DH compared fine-tuned InstructBLIP with retrieval-augmented variants and found that retrieval is useful only when retrieved examples are relevant [11]. The AUEB NLP Group investigated concept-aware multisynthesizers, MedCLIP-based reranking, and reinforcement learning objectives [12]. JJ-VMed explored LLaVA-style models and post-processing validation for automated medical image captions [21].

The proposed system follows the trend of using a large VLM, but keeps the captioning pipeline lightweight. Qwen3-VL-8B is not fine-tuned, and the pipeline does not use retrieval or train a learned reranker. Instead, the pipeline relies on instruction-following behavior and uses predicted CUI terms as soft clinical hints. The final output is then normalized by deterministic post-processing.

2.3. Biomedical image analysis pipelines

Recent biomedical image analysis studies have also demonstrated the value of robust engineering, preprocessing, and hybrid deep learning architectures. Nguyen-Tat et al. proposed a hybrid UNet, attention, and transformer approach for MRI brain tumor segmentation [22]. They also evaluated preprocessing and deep learning methods across multiple imaging modalities [23], while QMaxViT-Unet+ investigated query-based MaxViT-Unet and edge enhancement for scribble-supervised medical image segmentation [24]. Other related works include liver tumor segmentation using an engineering-based biomedical imaging framework [25] and knee osteoarthritis severity diagnosis with a GA-enhanced deep ensemble [26]. Although these studies address segmentation or diagnosis rather than caption generation, they motivate the engineering-oriented design of the system: modular processing, validation, and conservative post-processing for heterogeneous medical images.

3. Task and Data

3.1. Task definition

The ImageCLEFmedical Caption 2026 Caption Prediction task requires one generated caption for each test image. The submitted file must contain two columns, ID and Caption. Each caption should concisely describe the medical content of the corresponding image, including modality, anatomy, and visible findings when identifiable.

The official evaluation combines several complementary metrics. Textual and semantic relevance is measured with metrics such as BERTScore, which is based on Bidirectional Encoder Representations from Transformers (BERT) [27], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [28], semantic similarity, and Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) [29]. Biomedical grounding is assessed with metrics such as the Medical Concept Annotation Toolkit (MedCAT) [30] and image-text alignment. These metrics emphasize that high-quality medical captioning requires more than fluent text: the generated caption should also contain clinically appropriate concepts.

3.2. Available inputs and resources

The official test set contains 15,249 biomedical images. The ImageCLEFmedical Caption 2026 data are based on an updated and extended version of the ROCov2 dataset [14, 2]. In addition to the image input, the pipeline uses auxiliary predicted CUIs generated by running the public UIT_Worker Concept Detection notebook [31]. These CUIs are mapped to textual biomedical terms using a CUI metadata dictionary and then inserted into the captioning prompt as optional hints. Table 1 summarizes the main resources used by the system.

Table 1
Main data sources used by the UIT_HighDefinition system.

Data source	Role	Usage in the system
Test images	Visual input	Used by Qwen3-VL-8B for caption generation.
Predicted CUIs	Clinical hints	Converted into textual terms and inserted as soft prompt guidance.
CUI metadata dictionary	Concept mapping	Maps UMLS CUIs to readable biomedical terms.
Raw captions	Intermediate output	Normalized by rule-based post-processing.
Submission CSV	Official output	Two-column file containing ID and Caption.

The final predicted CUI file contains 15,249 rows. Among them, 115 images have no predicted CUI. The average number of predicted CUIs per image is 1.37, the median is 1, and the maximum is 7. Therefore, the maximum prompt budget of eight terms was never exceeded in the final test predictions.

4. System Overview and Methodology

Figure 1 shows the overall workflow. For each test image, externally generated auxiliary CUI predictions provide possible UMLS concepts. These CUIs are mapped to readable terms and inserted into a structured medical captioning prompt. Qwen3-VL-8B generates a raw caption, which is then normalized by a rule-based cleaner. The cleaned captions from all distributed workers are merged and validated to produce the final submission.

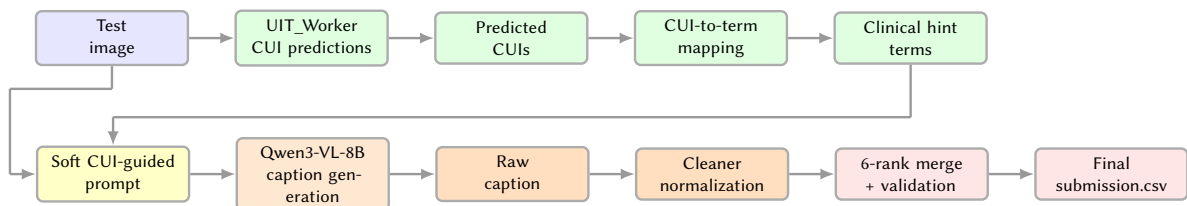


Figure 1: Overall CUI-guided Qwen3-VL medical image captioning pipeline. Externally generated predicted CUIs are mapped to readable biomedical terms and injected as soft prompt guidance before caption generation and rule-based post-processing.

4.1. Auxiliary CUI prediction

The auxiliary CUI predictions were generated by running the public UIT_Worker Concept Detection notebook for ImageCLEFmedical 2026 [31]. The Concept Detection training and inference code was not modified in this work. The external concept detector is a DenseNet121-based multi-label classifier trained for the Concept Detection subtask over 2,646 UMLS concepts. In the UIT_HighDefinition Caption Prediction pipeline, its output is used only as an auxiliary source of candidate UMLS CUIs for the official test images.

For transparency, the public UIT_Worker notebook uses RGB conversion, resizing to 224×224 , ImageNet normalization, and lightweight training augmentation for the underlying DenseNet121 concept detector. The training augmentation includes random shift, scale, and rotation with shift limit 0.05, scale limit 0.05, rotation limit 10 degrees, and probability 0.5, as well as random brightness and contrast adjustment with brightness limit 0.2, contrast limit 0.2, and probability 0.5. Validation and inference use only resizing, ImageNet normalization, and tensor conversion. At inference time, sigmoid probabilities are thresholded at $\tau = 0.60$, and labels above this threshold are decoded back into CUI strings.

For image i , the final textual hint set used by the captioning system is defined as:

$$T_i = \text{TopK}(\{m(c) \mid c \in C_i\}, K), \quad (1)$$

where C_i is the set of externally predicted CUIs for image i , $m(c)$ maps a CUI to a readable biomedical term, and $K = 8$. The CUI terms are not treated as ground-truth labels and are not forced into the generated caption. They are only optional clinical hints for the captioning prompt.

4.2. Soft CUI-guided prompt construction

The prompt is designed around three principles. First, the image remains the primary evidence source. Second, CUI terms are provided only as possible hints. Third, the caption should be concise, medically relevant, and visually grounded. In preliminary experiments, weaker prompts often produced generic captions, unsupported diagnoses, or verbose multi-sentence outputs. The final prompt therefore explicitly requests a one-sentence caption, encourages modality and anatomy identification, and warns the model not to force unsupported CUI terms into the output.

This soft-guidance design differs from hard concept injection. If a CUI is visually supported, the model may use it to improve biomedical specificity. If it is noisy or unsupported, the model is instructed to ignore it. This trade-off is important because predicted CUIs can be incomplete or incorrect. The exact prompt template used in the final submission is shown below.

You are generating a caption for a medical figure in a biomedical article.

Task:

Write one concise, clinically relevant English caption for the image.

Rules:

- Focus on the visible medical content.
- Mention modality when clear: CT, MRI, X-ray, ultrasound, histology, endoscopy, angiography, PET/CT, microscopy.
- Mention anatomical region and main visual finding if visible.
- Do not write "This image shows".
- Do not write "Caption:".
- Do not include uncertainty unless needed.
- Do not invent patient age, sex, diagnosis, treatment, or outcome if not visible.
- Output only one caption sentence.

Possible UMLS/CUI clinical hints: {cui_terms}

Use these hints only if they are visually supported by the image.
Do not force unsupported diseases or findings.

Final caption:

4.3. Qwen3-VL caption generation

The caption generator is Qwen3-VL-8B-Instruct [13]. The model is used in inference-only mode without fine-tuning. It was selected because it provides a practical balance between multimodal instruction-following capability and feasibility under Kaggle GPU constraints. To fit the model into the available environment, the implementation uses 4-bit quantization and deterministic generation. The main generation settings include `do_sample=False` and `max_new_tokens=80`. Each image is processed independently together with the corresponding prompt and optional CUI terms.

4.4. Caption post-processing

Raw VLM outputs may include boilerplate prefixes, unnatural modality expressions, overly long descriptions, or invalid endings. A conservative rule-based cleaner is therefore applied. The cleaner removes prefixes such as “Caption:” and “This image shows”, normalizes modality expressions such as “Magnetic Resonance Imaging” to “MRI” and “Computed Tomography” to “CT”, trims overly long captions, repairs selected truncated endings, filters generic captions, and enforces final punctuation. The cleaner is deliberately conservative: it improves formatting and linguistic consistency but does not attempt to verify medical factual correctness.

4.5. Distributed inference and submission assembly

Processing 15,249 images with a large VLM is time-consuming. Inference is therefore distributed across six GPU workers. The test IDs are sorted and deterministically assigned to ranks using modulo sharding:

$$r(i) = i \bmod 6. \quad (2)$$

The expected rank sizes are 2,542 images for ranks 0–2 and 2,541 images for ranks 3–5. Each rank writes a CSV file and saves progress after every image. The final merge script validates the row count, duplicate IDs, empty captions, and missing/extra IDs before creating the submission ZIP. Figure 2 illustrates this workflow.

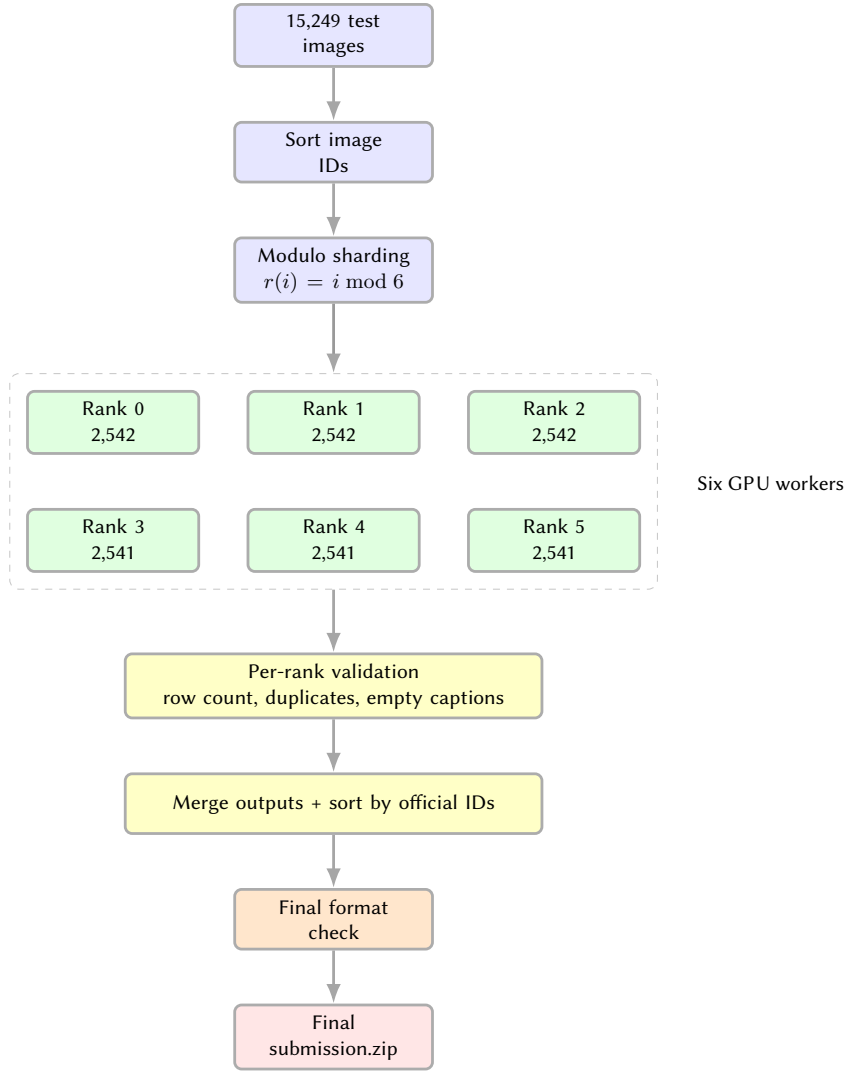


Figure 2: Distributed six-rank inference and submission assembly workflow. The test set is deterministically sharded by modulo assignment, validated rank by rank, then merged and checked to produce the final submission.

5. Experimental Setup and Results

5.1. Implementation details

Table 2 summarizes the main implementation settings. The captioning stage uses Qwen3-VL-8B-Instruct with 4-bit quantization. Auxiliary CUI hints are derived from externally generated UIT_Worker Concept Detection outputs. The final cleaner limits captions to at most 55 words after post-processing.

Table 2

Implementation details of the submitted system.

Parameter	Value
Captioning backbone	Qwen3-VL-8B-Instruct
Captioning mode	Inference only, no fine-tuning
Quantization	4-bit
Generation setting	Deterministic, do_sample=False
Maximum new tokens	80
Auxiliary CUI source	Public UIT_Worker Concept Detection outputs
Source CUI label space	2,646
CUI threshold	0.60
Maximum CUI terms in prompt	8
Maximum caption length after cleaning	55 words
Test images	15,249
Distributed workers	6 ranks across 3 machines
Checkpoint interval	After every image

5.2. Official final result

Table 3 reports the official scores of the final submission. The final run used Qwen3-VL-8B with CUI-guided prompting and rule-based cleaning. It achieved an overall score of 0.3193 and ranked seventh on the official Caption Prediction Test 2026 leaderboard. In the official Caption Prediction Test 2026 leaderboard, the final submission was reported under the final team name `UIT_HighDefinition` with run ID 980; the original leaderboard username was `tanduy2702`. Table 4 provides the official leaderboard context for the top-ranked submissions and the final run.

Table 3Official score of the final `UIT_HighDefinition` submission.

System	Overall	Relevancy	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
Qwen3-VL-8B + CUI prompt + cleaner	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379

Table 4

Official leaderboard context for Caption Prediction Test 2026. The table reports ranks 1–8 from the final official leaderboard, using the final team names requested by the organizers.

Rank	Team	Run ID	Overall	Relevancy	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
1	AlstatLab	937	0.3755	0.5786	0.1724	0.6250	0.3234	1.0104	0.3555	0.2160	0.1287
2	UFPE-Essex-UNED	977	0.3603	0.5527	0.1679	0.6017	0.2725	0.9902	0.3463	0.1822	0.1537
3	dsgt_caption	990	0.3571	0.5365	0.1777	0.6042	0.2809	0.9381	0.3229	0.1943	0.1610
4	AUEB/DMST	985	0.3516	0.5288	0.1743	0.6130	0.2728	0.9113	0.3183	0.1812	0.1674
5	csmorgan	936	0.3458	0.5377	0.1539	0.6075	0.2725	0.9408	0.3299	0.1751	0.1327
6	gfreitas	840	0.3343	0.5201	0.1484	0.6013	0.2226	0.9273	0.3293	0.1563	0.1406
7	UIT_HighDefinition	980	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379
8	NUSTPB	774	0.3065	0.4705	0.1426	0.5465	0.2224	0.7787	0.3342	0.1443	0.1408

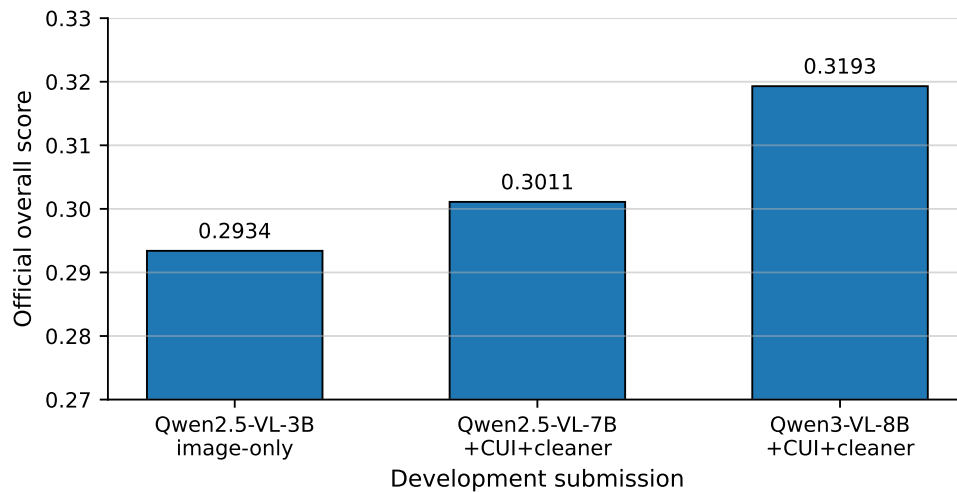
5.3. Development submissions

Table 5 summarizes the evolution of the main development submissions. The initial Qwen2.5-VL-3B image-only baseline achieved an overall score of 0.2934. A Qwen2.5-VL-7B variant with soft CUI prompting and cleaner improved the score to 0.3011. The final Qwen3-VL-8B system obtained the best overall score, 0.3193. Figure 3 visualizes this improvement.

Table 5

Development submissions and official feedback scores. These are development variants rather than a controlled ablation, because several components changed between runs.

System variant	Overall	Relevancy	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
Qwen2.5-VL-3B image-only	0.2934	0.4764	0.1103	0.5902	0.2289	0.7622	0.3243	0.1106	0.1100
Qwen2.5-VL-7B + CUI + cleaner	0.3011	0.4880	0.1143	0.5865	0.2267	0.8153	0.3235	0.1239	0.1046
Qwen3-VL-8B + CUI + cleaner	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379

**Figure 3:** Development submission evolution based on official feedback scores.

Compared with the image-only baseline, the final system improved the overall score by 0.0259, similarity from 0.7622 to 0.8496, MedCAT from 0.1106 to 0.1397, and alignment from 0.1100 to 0.1379. ROUGE slightly decreased from 0.2289 to 0.2243, suggesting that the final system improved semantic and clinical relevance without necessarily increasing lexical overlap with reference captions.

5.4. Post-processing validation

Table 6 reports the main cleaner validation statistics. For the Qwen2.5-VL-7B cleaned submission, the cleaner changed 1,686 rows, reduced bad modality expressions from 1,507 to 0, and reduced truncated endings from 18 to 0. For the final Qwen3 system, the cleaner reduced detected bad modality expressions from 461 to 0. Table 7 provides examples of modality normalization performed by the cleaner.

Table 6

Post-processing validation statistics. The checks verify formatting and surface-level consistency, not clinical factual correctness.

Check	Before cleaning	After cleaning
Bad modality terms in Qwen2.5-VL-7B	1,507	0
Truncated endings in Qwen2.5-VL-7B	18	0
Bad modality expressions in Qwen3	461	0

Table 7

Examples of modality normalization performed by the cleaner.

Raw output fragment	Cleaned output fragment	Rule
Magnetic Resonance Imaging of the brain revealing multiple hyperintense lesions	MRI of the brain revealing multiple hyperintense lesions	Modality normalization
Lateral plain x-ray of the left shoulder demonstrating a humeral shaft fracture	Lateral X-ray of the left shoulder demonstrating a humeral shaft fracture	Modality normalization
Abdominal plain radiograph demonstrating dilated bowel loops	Abdominal X-ray demonstrating dilated bowel loops	Modality normalization

5.5. CUI distribution and final submission validation

Figure 4 shows the distribution of predicted CUI hints per image. Most images contain one or two CUI hints, while 115 images contain no predicted CUIs. This supports the use of soft CUI prompting without overloading the input prompt.

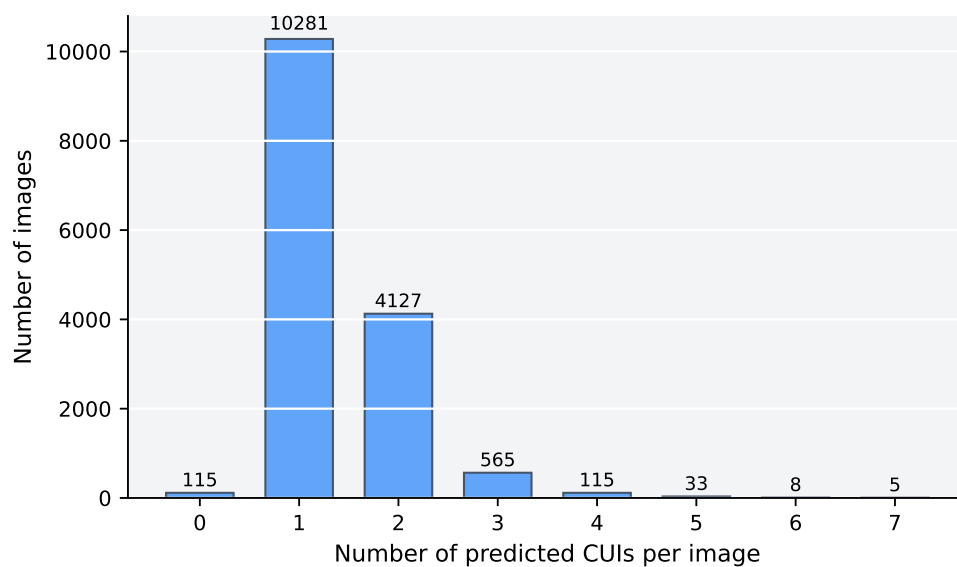


Figure 4: Distribution of predicted CUI hints per test image.

The final submission passed internal validation: 15,249 rows, no duplicate IDs, no empty captions, the required two-column header ID, Caption, and 13,666 unique final captions. Table 8 summarizes these checks.

Table 8

Final submission validation.

Validation item	Value
Expected test images	15,249
Submission rows	15,249
Duplicate IDs	0
Empty captions	0
Unique captions	13,666
Images without predicted CUIs	115
Mean predicted CUIs per image	1.37
Median predicted CUIs per image	1
Maximum predicted CUIs per image	7

6. Discussion

6.1. Why soft CUI guidance helped

The development submissions suggest that concept guidance improved semantic relevance and biomedical concept coverage. Moving from the Qwen2.5-VL-3B image-only baseline to stronger CUI-guided systems increased the overall score, similarity, MedCAT, and final alignment. This indicates that predicted biomedical concepts can help a VLM focus on medically relevant terminology. However, the Qwen2.5-VL-7B CUI variant slightly reduced BERT, ROUGE, and alignment compared with the image-only baseline, showing that concept guidance can introduce trade-offs. For this reason, CUIs are used as optional hints rather than required words.

6.2. Semantic similarity versus clinical grounding

The final system achieved high semantic similarity but lower MedCAT and alignment scores. This gap suggests that the model often produced fluent and semantically plausible captions, but exact biomedical concept grounding remained difficult. A caption may correctly describe the modality and general anatomy while missing the exact disease, anatomical detail, or clinical concept expected by MedCAT. This is a common challenge for medical captioning: language-level relevance is easier to achieve than precise clinical concept matching.

6.3. Role of the cleaner

The cleaner improved formatting consistency and removed common VLM artifacts. Its strongest effect was modality normalization, where phrases such as “Magnetic Resonance Imaging” and “plain radiograph” were mapped to shorter standardized terms such as “MRI” and “X-ray”. This makes captions more concise and consistent with medical caption style. However, the cleaner does not verify visual evidence or factual correctness. Therefore, it should be viewed as a submission-stability component rather than a clinical reasoning module.

6.4. Failure observations and future reranking

During development, several recurring failure modes were observed: generic captions, wrong or vague modality, unsupported disease mentions, missing clinical concepts, histology/microscopy ambiguity, and multi-panel figure confusion. Diagnostic scoring components such as coverage estimates, hallucination penalties, annotation penalties, and generic-template penalties were also explored during development. These components were used for debugging and analysis, not as the final learned reranker. A promising future direction is to generate multiple candidate captions and select among them using a medical reranking function that jointly considers concept coverage, visual grounding, and hallucination risk.

6.5. Reproducibility considerations

The final submission was assembled from six distributed rank CSV files. The public repository provides the prompt template, post-processing logic, merge and validation scripts, notebooks, environment notes, and result summaries. It does not redistribute ImageCLEFmedical dataset images, ground-truth annotations, final submission files, full generated captions, private logs, machine rank outputs, or the full leaderboard CSV. Exact recreation of the submitted file would therefore require authorized access to the official data and the private intermediate outputs used during submission assembly. The merge script was designed to stop if any rank file was missing, incomplete, duplicated, or empty.

6.6. Limitations

The submitted system has several limitations. First, Qwen3-VL-8B was not fine-tuned on the development captions, so the model may not fully match the reference caption style. Second, the predicted CUIs are noisy auxiliary signals and may be incomplete. Third, the rule-based cleaner cannot verify whether a generated statement is visually or clinically correct. Fourth, the system does not segment multi-panel figures or perform region-level grounding. Finally, no learned medical reranker is used in the final submission.

7. Conclusion

This paper presented the UIT_HighDefinition system for ImageCLEFmedical Caption 2026 Caption Prediction. The system combines Qwen3-VL-8B-Instruct with soft CUI-guided prompting based on externally generated CUI predictions, rule-based caption cleaning, and six-rank distributed inference. The final official submission achieved an overall score of 0.3193 and ranked seventh on the leaderboard. Development results showed consistent improvement from an image-only baseline to the final CUI-guided Qwen3-VL system. The remaining gap between semantic similarity and clinical grounding suggests that future work should focus on stronger concept validation, candidate reranking, visual grounding, and possibly domain-specific fine-tuning.

Acknowledgments

We thank the ImageCLEFmedical Caption organizers for providing the benchmark, evaluation platform, and official feedback.

This research was funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number CS4-2026-01.

Code Availability

The source code, prompt template, post-processing scripts, merge and validation scripts, environment notes, and reproducibility documentation are available at: https://github.com/quytsua/UIT_HighDefinition_ImageCLEFmedical_Caption_2026. The repository does not redistribute ImageCLEFmedical dataset images, ground-truth annotations, private rank outputs, or the final submitted caption file.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) for the following GenAI contribution roles: Grammar and spelling check, Paraphrase and reword, Improve writing style, Formatting assistance, and Peer review simulation for manuscript and repository-readiness review. ChatGPT was not used to generate the experimental results, produce the submitted captions, access private benchmark

data, or replace the authors' scientific interpretation. All result tables and figures were derived from official scores, validation logs, or author-verified experimental records. After using this tool, the authors reviewed, edited, and verified the content as needed. The final outputs, results, claims, and manuscript were approved by the authors, who take full responsibility for the content of this publication.

References

- [1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [2] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryilmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [3] J. Rückert, L. Bloch, A. Idrissi-Yaghir, R. Brüngel, H. Schäfer, S. Koitka, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – caption prediction and concept detection, in: *CLEF 2023 Working Notes, volume 3497 of CEUR Workshop Proceedings, 2023*. URL: <https://ceur-ws.org/Vol-3497/paper-108.pdf>.
- [4] J. Rückert, H. Damm, A. Idrissi-Yaghir, R. Brüngel, L. Bloch, H. Schäfer, S. Koitka, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – caption prediction and concept detection, in: *CLEF 2024 Working Notes, volume 3740 of CEUR Workshop Proceedings, 2024*. URL: <https://ceur-ws.org/Vol-3740/paper-132.pdf>.
- [5] 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. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation, in: *CLEF 2025 Working Notes, volume 4038 of CEUR Workshop Proceedings, 2025*. URL: https://ceur-ws.org/Vol-4038/paper_170.pdf.
- [6] A. Nicolson, J. Dowling, B. Koopman, A concise model for medical image captioning, in: *CLEF 2023 Working Notes, volume 3497 of CEUR Workshop Proceedings, 2023*. URL: https://github.com/aehrc/imageclefmedical_caption_23.
- [7] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, J. S. Cardoso, Detecting concepts and generating captions from medical images: Contributions of the VCMi team to ImageCLEFmedical caption 2023, in: *CLEF 2023 Working Notes, volume 3497 of CEUR Workshop Proceedings, 2023*. URL: <https://ceur-ws.org/Vol-3497/paper-137.pdf>.
- [8] T. V. Phan, T. K. Nguyen, Q. A. D. D. Hoang, Q. T. Phan, T. B. Nguyen-Tat, UIT-2Q2T at ImageCLEFmedical 2024 caption: Multimodal medical image captioning using bootstrapping language-image pre-training, in: *CLEF 2024 Working Notes, volume 3740 of CEUR Workshop Proceedings, 2024*. URL: <https://ceur-ws.org/Vol-3740/paper-159.pdf>.
- [9] Q. V. Nguyen, H. Q. Pham, D. Q. Tran, T. K.-B. Nguyen, N.-H. Nguyen-Dang, T. B. Nguyen-Tat, UIT-DarkCow team at ImageCLEFmedical caption 2024: Diagnostic captioning for radiology

- images efficiency with transformer models, in: CLEF 2024 Working Notes, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: <https://ceur-ws.org/Vol-3740/paper-158.pdf>.
- [10] N. N.-Y. Nguyen, H. L. Tu, P. D. Nguyen, T. N. Do, T. M. Thai, T. B. Nguyen-Tat, DS@BioMed at ImageCLEFmedical caption 2024: Enhanced attention mechanisms in medical caption generation through concept detection integration, in: CLEF 2024 Working Notes, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: <https://ceur-ws.org/Vol-3740/paper-157.pdf>.
 - [11] J. He, S. Ferdowsi, W. Feng, F. Alves, A. Platon, D. Teodoro, DS4DH group at ImageCLEFmedical caption 2025, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025.
 - [12] A. Chatzipapadopoulou, I. Pantelidis, F. Charalampakos, M. Samprovalaki, G. Moschovis, P. Kaliosis, K. V. Dalakleidi, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical caption 2025, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025.
 - [13] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, K. Zhu, Qwen3-VL technical report, arXiv preprint arXiv:2511.21631 (2025). URL: <https://arxiv.org/abs/2511.21631>.
 - [14] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCov2: Radiology Objects in Context version 2, an updated multimodal image dataset, *Scientific Data* 11 (2024) 688. doi:10.1038/s41597-024-03496-6.
 - [15] H. Kauschke, K. Bogomasov, S. Conrad, Predicting captions and detecting concepts for medical images: Contributions of the DBS-HHU team to ImageCLEFmedical caption 2024, in: CLEF 2024 Working Notes, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: <https://ceur-ws.org/Vol-3740/paper-153.pdf>.
 - [16] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Research* 32 (2004) D267–D270. doi:10.1093/nar/gkh061.
 - [17] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: *International Conference on Machine Learning*, 2022, pp. 12888–12900.
 - [18] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: *International Conference on Machine Learning*, 2023, pp. 19730–19742.
 - [19] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. C. H. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in: *Advances in Neural Information Processing Systems*, volume 36, 2023.
 - [20] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, R. Valencia-García, UMUTeam at ImageCLEF 2025, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 2553–2563.
 - [21] J. Angulo, J. Aguilar, JJ-VMed: A framework for automated concepts, captions and explainability of medical image, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_179.pdf.
 - [22] T. B. Nguyen-Tat, T.-Q. T. Nguyen, H.-N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in MRI images: A hybrid approach using UNet, attention mechanisms, and transformers, *Egyptian Informatics Journal* 27 (2024) 100528. doi:10.1016/j.eij.2024.100528.
 - [23] T. B. Nguyen-Tat, Q. H. Tran, T. N. Pham, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, *Alexandria Engineering Journal* 119 (2025) 558–586. doi:10.1016/j.aej.2025.01.090.
 - [24] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images, *Computers in Biology and Medicine* 187 (2025) 109762. doi:10.1016/j.combiomed.2025.109762.
 - [25] V. Pham, H. Nguyen, B. Pham, T. Nguyen, H. Nguyen, Robust engineering-based unified biomedical

- imaging framework for liver tumor segmentation, *Current Medical Imaging* 19 (2022) 37–45. doi:10.2174/1573405617666210804151024.
- [26] T. B. Nguyen-Tat, T.-P. Nguyen-Duong, Optimizing knee osteoarthritis severity diagnostics: A GA-enhanced deep ensemble approach in medical imaging, *Ain Shams Engineering Journal* 16 (2025) 103524. doi:10.1016/j.asej.2025.103524.
 - [27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: *International Conference on Learning Representations*, 2020.
 - [28] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, 2004, pp. 74–81.
 - [29] T. Sellam, D. Das, A. P. Parikh, BLEURT: Learning robust metrics for text generation, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7881–7892.
 - [30] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. B. Dobson, Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit, *Artificial Intelligence in Medicine* 117 (2021) 102083. doi:10.1016/j.artmed.2021.102083.
 - [31] T. Q. To, D. C. Pham, N. H. Bui, T. B. Nguyen-Tat, ImageCLEFmedical 2026 UIT_Worker source code for concept detection, <https://github.com/nhanbui246/ImageCLEFmedical-2026-UIT-Worker>, 2026. Public repository for the ImageCLEFmedical 2026 Concept Detection subtask. Accessed: 2026-06-26.