# NUSTPB at ImageCLEFmed Caption 2026: Medical Image Captioning with QLoRA-Adapted LLaVA-1.5*

ImageCLEFmed Caption at CLEF 2026

Mihail Cristian Tatomir[1,*], Dan Cristian Stanciu[1] and Liviu Daniel Ştefan[1]

[1]National University of Science and Technology Politehnica Bucharest, Romania

**Abstract**

In this paper, we present the NUSTPB submission to ImageCLEFmedical Caption 2026, which targets concept detection and caption prediction on real and synthetic radiology images. Our system adapts LLaVA-1.5-7B through Quantized Low-Rank Adaptation (QLoRA) and a unified instruction prompt that produces modality, UMLS Concept Unified Identifiers (CUIs), and a descriptions decoder in a single forward pass. On real data, our best run fro caption prediction reaches an overall caption score of 0.3065, while the best run for concept detection achieves a F1 score of 0.5594. On synthetic data, we achieve a caption score of 0.3095, with concept F1 dropping to 0.1048, which exposes an asymmetric domain gap between the language model and the visual encoder.

**Keywords**

Medical image captioning, Concept detection, LLaVA, QLoRA, Vision-language models.

## 1. Introduction

Radiology images contain clinically meaningful visual information that is commonly communicated through captions, structured terminology, and radiological reports. Producing these descriptions manually requires expert knowledge and remains a labor intensive step in clinical documentation, indexing, and biomedical literature search. Automatic medical image captioning addresses this need by mapping visual evidence to concise textual descriptions, while medical concept detection provides a complementary structured representation through Unified Medical Language System Concept Unique Identifiers (UMLS CUIs). These two outputs are closely related since a clinically useful caption should preserve the relevant modality, anatomy, visual findings, and medical concepts without introducing unsupported information.

ImageCLEF has provided a recurring benchmark environment for multimodal retrieval, classification, and generation since 2003, with medical tasks forming one of its most established directions [1]. Within this broader evaluation campaign, ImageCLEFmedical Caption focuses on radiology image understanding through concept detection and caption prediction [2]. The 2026 edition is the 10th edition of the Caption task and extends the benchmark with synthetic concept detection and synthetic caption prediction, in addition to the standard real data subtasks and the explainability task. This extension is particularly relevant for assessing whether models trained on real radiology data can generalize to images with synthetic visual characteristics, while the task rules require the standard and synthetic datasets to remain separated during system development.

The ImageCLEFmedical Caption series has progressively refined both the data and the evaluation protocol. The 2017 and 2018 editions introduced caption prediction and concept extraction for biomedical images, while the 2019 and 2020 editions narrowed the focus to radiology concept prediction and incorporated modality information [3, 4, 5, 6]. In 2021, both concept detection and caption prediction

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

*Corresponding author.

mihail.tatomir@stud.etti.upb.ro (M. C. Tatomir); dan.stanciu1203@upb.ro (D. C. Stanciu); liviu_daniel.stefan@upb.ro (L. D. Ştefan)

https://lstefan.aimultimedialab.ro/ (L. D. Ştefan)

https://orcid.org/0009-0008-4561-4328 (D. C. Stanciu); 0000-0001-9174-3923 (L. D. Ştefan)

were run again with a stronger emphasis on medically relevant annotations, and the following editions continued to improve dataset quality, concept filtering, and evaluation [7, 8, 9]. The 2023 edition adopted BERTScore as the primary caption metric, reflecting the need to evaluate semantic similarity beyond surface overlap. The 2024 edition introduced additional caption metrics and an optional explainability extension, while the 2025 edition promoted explainability to an official subtask and used a composite caption score that combined relevance and factuality dimensions [10, 11].

Recent ImageCLEFmedical Caption results show a methodological transition from conventional encoder decoder models and convolutional multi label classifiers toward large Vision Language Models adapted to the medical domain. In 2024, strong captioning systems used medical vision language foundation models, multimodal large language models, and modality aware adaptation, while concept detection remained highly competitive with optimized convolutional ensembles [12]. In 2025, the caption prediction subtask showed a clearer shift toward fine tuning Vision Language Models such as BLIP, InstructBLIP, LLaVA based systems, and modular multimodal pipelines, although factuality oriented metrics remained difficult for all participants [13]. These observations motivate approaches that adapt powerful general multimodal models to radiology while controlling computational cost and generation behavior.

In this paper, we describe the NUSTPB submission to ImageCLEFmedical Caption 2026. Our system adapts LLaVA 1.5 7B using QLoRA, enabling parameter efficient specialization of a general instruction following Vision Language Model for medical image captioning and concept prediction [14, 15]. Instead of training separate models for the two main outputs, we use a unified instruction format that asks the model to identify the imaging modality, list the relevant CUIs, and describe the findings in a single response. This formulation aligns concept prediction and caption generation within one generative pipeline and allows the same adapted model to be evaluated on both standard and synthetic subtasks. We further study decoding and post processing choices, including deterministic generation, beam search, repetition control, and sentence level cleaning, since medical captioning requires concise and factual outputs rather than creative variation. Through this design, the paper contributes a compact evaluation of parameter efficient Vision Language Model adaptation under the real and synthetic data conditions introduced in the 2026 task.

## 2. Dataset and Task

The dataset basis for the recent editions is ROCOv2, an updated version of Radiology Objects in Context that provides radiology images from the PubMed Central Open Access subset together with captions and associated medical concepts [16]. In the 2026 task, ROCOv2 remains the development data source for both concept detection and caption prediction, and the concepts are derived from a reduced UMLS 2022AB release after semantic and frequency based filtering [2]. This setting encourages systems that can jointly exploit visual information, medical terminology, and caption style, while avoiding overgeneration of concepts that are unlikely to be visually grounded.

For the concept detection subtask, the input is a medical images, and the expected output is a set of UMLS CUIs that describe the visible medical content. Performance is measured using F1 scores between the predicted and reference concept sets. For the standard concept detection task, a secondary score is also computed on a manually curated subset of concepts. For the synthetic concept detection task, only the primary F1 score is reported.

For the caption prediction subtask, the input is again a medical image, while the expected output is a coherent caption describing the image content. The generated caption should capture the imaging modality, anatomical structures, visible findings, and clinically relevant abnormalities. In contrast to concept detection, which evaluates set coverage, caption prediction evaluates both the relevance and factuality of the generated text. The official ranking is based on an average score over six metrics. Relevance is assessed using image-caption similarity, BERTScore recall with inverse document frequency weighting, ROUGE-1, and BLEURT. Factuality is assessed using UMLS Concept F1 and AlignScore.

Before computing BERTScore, ROUGE, and BLEURT, captions are lowercased, numbers are replaced with a generic number token, punctuation is removed, and each caption is treated as a single sentence.

# 3. Methodology

## 3.1. Base Model

We adapt the LLaVA-1.5-7B checkpoint (llava-hf/llava-1.5-7b-hf)[14]. The model combines a CLIP ViT-L/14 visual encoder operating at $336 \times 336$ input resolution with the Vicuna-7B language model. Visual features are mapped into the language model's embedding space through a multilayer perceptron projection. The pretraining corpus is dominated by non-medical image-text pairs, so the model arrives with general visual grounding and fluent instruction following but without specialized radiology vocabulary or CUI awareness. Our adaptation therefore targets domain-specific terminology and visual-to-clinical association rather than basic language or visual capability.

## 3.2. QLoRA Adaptation

Full fine-tuning of the 7B-parameter backbone is memory-prohibitive on our hardware and risks degrading the model's general linguistic capabilities. We use QLoRA [15] to combine 4-bit quantization of the frozen base model (NF4 format with double quantization) with trainable low-rank adapters. Forward and backward computations are performed in bfloat16 to retain sufficient numerical precision for the fine-grained visual cues present in radiographs and CT scans. Gradient checkpointing and Scaled Dot-Product Attention (SDPA) further reduce activation memory.

The LoRA adapter is configured with rank $r = 64$ and scaling factor $\alpha = 128$. We selected a relatively high rank ($r = 64$) as a design choice to provide greater adaptation capacity for the radiology domain. We use a LoRA dropout of 0.05 and AdamW weight decay of 0.01 to regularize the adapter. Training proceeds in two stages with different attention coverage:

**Stage 1 (domain adaptation).** Learning rate $2 \times 10^{-5}$, with LoRA applied to all four self-attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`). This stage incorporates clinical vocabulary and CUI structure across the full attention pathway.

**Stage 2 (refinement).** Learning rate reduced to $5 \times 10^{-6}$ and LoRA restricted to `q_proj` and `v_proj`, resuming from the Stage 1 checkpoint. The narrower coverage and lower learning rate consolidate clinical associations without disturbing the broader fluency acquired in Stage 1.

The per-device batch size is 1 due to VRAM constraints, with gradient accumulation over 32 steps to provide a stable optimization signal. To ensure training stability, memory constraints mandated a batch size of 1 image per iteration, which we compensated for by using gradient accumulation over 32 steps. This ensured that the model analyzed a representative volume of data before updating its weights, thereby eliminating statistical noise. Additionally, we implemented a loss masking strategy by setting the labels corresponding to the prompt instructions to a negative constant, so that the training loss was computed only over the generated medical descriptions rather than the prompt tokens.

## 3.3. Prompt Design

LLaVA-1.5 was instruction-tuned in a conversational format with `USER:` and `ASSISTANT:` role markers, which we retain during adaptation to avoid distribution shift in the conversational scaffolding. The prompt used for both training and inference is:

```
USER: <image>
Identify the modality, list the medical CUIs and describe the findings.
ASSISTANT:
```

The <image> token marks the insertion point for the projected visual tokens produced by the CLIP encoder. The single-prompt design elicits all three target fields (modality, CUI list, findings) in one decoding pass and removes the need to maintain separate training pipelines or models for the concept detection and caption prediction subtasks.

During training we apply loss masking: tokens belonging to the prompt are assigned label $-100$ so that the cross-entropy is computed only over the target completion, with the model receiving no learning signal for reproducing the fixed instruction.

## 3.4. Submission Strategy

For our participation in the competition tasks, we apply the same adapted checkpoint across all four target subtasks.All training and evaluation procedures were conducted using a Google Colab environment equipped with an NVIDIA A100 GPU. In accordance with the rule that the standard and synthetic datasets must not be mixed, the checkpoint is trained only on the real ROCOv2 training data, using an 80-20 train-test random split. Its application to the synthetic subtasks therefore constitutes a zero-shot transfer evaluation. Two checkpoints from the training run were considered for submission. The first one, `checkpoint-1500`, which represents the model after 1500 training steps (approximately 1 full epoch),is used in most runs. The second one, `checkpoint-1300`, representing the model after 1300 training steps (approximately 0.85 epochs), is used in the final synthetic submission to probe whether further training improves robustness to the synthetic visual distribution. The five submissions allowed per subtask varied only the inference-time configuration described in Section 3.5. The underlying model was held constant to observe how the model reacts to different constraints in distinct environments.

## 3.5. Decoding and Post-Processing Configurations

This section describes each submitted configuration in the order in which it was submitted, separately for the real and the synthetic subtasks. Each description states the parameter values, the motivation for the configuration, and the property of the system that it is designed to probe. The numerical results obtained by each configuration are reported in Section 4.

### 3.5.1. Real-Data Submissions

**Run 1: Greedy decoding.** The first real-data submission uses greedy decoding with `do_sample=False` and `num_beams=1`. Greedy decoding selects the highest-probability token at each step and provides a deterministic baseline that reflects the raw output of the adapted model under the simplest decoding policy. We use it to establish a clean reference against which the gains of more elaborate strategies are measured. We expect this configuration to be a strong concept detector, because the dominant CUIs receive high posterior probability under the adapted weights. We do not expect it to be a competitive captioner, because greedy generation tends to repeat token spans in longer sequences and frequently fails to terminate on a complete sentence within the token budget.

**Run 2: Beam-2 decoding with sentence-level deduplication.** The second real-data submission combines beam search at width two with a post-processing step that splits the generated text into sentences, removes semantic duplicates, and rejoins the remaining sentences. The beam search component is intended to mitigate the repetition observed under greedy decoding by exploring multiple sequence trajectories before committing to the most probable complete sequence. The deduplication component targets the concise structure of the reference captions by removing redundant sentences that inflate the n-gram count without contributing new clinical content. Both components operate independently of the CUI list, which is unchanged by the post-processing step. We expect this configuration to retain the strong concept performance of Run 1 while delivering a substantial gain on the caption metrics.

**Run 3: Beam-2 decoding without post-processing.** The third real-data submission applies the same beam search as Run 2 but omits the deduplication step. The configuration isolates the contribution of post-processing by enabling a direct comparison with Run 2. Any difference between the two runs is attributable to the deduplication step, since the underlying generations are otherwise produced under identical decoding parameters.

**Run 4: Stochastic sampling.** The fourth real-data submission enables stochastic sampling with `do_sample=True` and temperature $T = 1.2$ (higher than the deterministic setting used in other runs).. Sampling flattens the next-token distribution and admits lower-probability tokens at each step. The configuration tests the hypothesis that increased lexical diversity improves the relevance metrics over deterministic decoding. The same mechanism, however, raises the probability of emitting tokens that are not visually grounded. We therefore expect a trade-off between diversity gains on the relevance side and factuality losses on the concept side, and we use this run to determine on which side the trade-off resolves in the medical domain.

**Run 5: Constrained beam-4 decoding with strict truncation.** The final real-data submission returns to deterministic decoding with `num_beams=4` and `no_repeat_ngram_size=2`. The wider beam explores more candidate sequences than Run 2, and the n-gram constraint prevents any 2-gram from appearing more than once in the generated sequence. The output is then post-processed by truncating at the last complete sentence boundary and capping the result at two unique sentences derived from the first 25 characters. The configuration targets dense and grammatically complete captions. It accepts the risk that valid material emitted late in the generated sequence is discarded by the strict truncation, which may affect concept recall when the model emits secondary CUIs at the end of the response.

### 3.5.2. Synthetic-Data Submissions

**Run 1: Beam-3 decoding with repetition penalty.** The first synthetic submission uses beam search at width three together with `repetition_penalty=1.2`. The wider beam and the explicit repetition penalty are intended to encourage commitment to specific clinical content under the artificial visual distribution, where preliminary inspection of the development set indicated that the CUI predictions were more diffuse than on the real data. We submitted two slots with this configuration to obtain a basic estimate of submission-level variance. Both submissions produced identical scores on both subtasks and are reported together as Run 1 in Section 4.

**Run 2: Confidence-threshold filtering on CUI predictions.** The second synthetic submission retains the beam-3 decoder of Run 1 and adds a probability threshold that filters low-confidence CUI predictions. The threshold was tuned on the development set. The configuration tests the hypothesis that the synthetic-domain shift inflates the rate of spurious low-confidence CUIs without affecting the rate of correct high-confidence CUIs. If the hypothesis holds, the filter should improve concept F1 substantially while leaving the surface fluency of the captions essentially unchanged. The caption-level metrics are therefore expected to drop only by the amount attributable to removing the discarded concepts from the caption text.

**Run 3: Length-constrained generation.** The third synthetic submission applies `max_new_tokens=90` and `length_penalty=0.8` to enforce caption density. The hypothesis is that the synthetic images induce verbose explanations of visual artifacts that dilute the reference-aligned content of the captions. The length constraint is therefore expected to improve the surface-overlap metrics by removing low-value tokens from the end of the generated sequence. The same mechanism, however, may also remove longer-tail CUIs that the model emits late in the response, in which case concept F1 would deteriorate.

**Table 1**
Caption prediction results on real data. Overall is the official average over the six caption metrics; higher is better for all columns. Runs are ordered by Overall. The "–" row indicates that Run 1 was submitted to concept detection only.

| Run | ID | Overall | BERT | ROUGE | Similarity | BLEURT | UMLS Concept F1 | Align |
|---|---|---|---|---|---|---|---|---|
| 2 | 774 | 0.3065 | 0.5465 | 0.2224 | 0.7787 | 0.3342 | 0.1443 | 0.1408 |
| 3 | 775 | 0.3011 | 0.4999 | 0.1862 | 0.7845 | 0.3404 | 0.1380 | 0.1610 |
| 5 | 956 | 0.2859 | 0.5669 | 0.2217 | 0.7386 | 0.3164 | 0.1221 | 0.0999 |
| 4 | 855 | 0.2797 | 0.5309 | 0.1876 | 0.7718 | 0.3065 | 0.1080 | 0.1123 |
| 1 | – | – | – | – | – | – | – | – |

**Table 2**
Concept detection results on real data. F1 is computed between predicted and reference CUI sets. Secondary F1 is computed on a manually curated subset of concepts. Runs are ordered by F1; higher is better for both columns.

| Run | ID | F1 | Secondary |
|---|---|---|---|
| 1 | 765 | 0.5594 | 0.9465 |
| 2 | 773 | 0.5540 | 0.9364 |
| 3 | 776 | 0.5540 | 0.9364 |
| 4 | 854 | 0.3407 | 0.5371 |
| 5 | 955 | 0.3084 | 0.4900 |

**Run 4: Extended-training checkpoint.** The final synthetic submission substitutes `checkpoint-1500` for `checkpoint-1300` and applies the strict-truncation post-processing of Run 5 in the real-data block. The configuration probes whether additional training on real ROCOv2 data improves transfer to the synthetic distribution. Two outcomes are possible. If the additional training steps improve the visual encoder's generalization, the synthetic concept F1 should rise. If the additional steps overfit the adapter to the real visual distribution, the synthetic results may instead worsen.

## 4. Results

We report results separately for the real and the synthetic subtasks. Each run is identified by the number assigned to it in Section 3.5, and the corresponding CodaBench submission identifiers appear in the result tables. For each subtask we present the metric tables and then discuss the runs in numerical order, comparing the observed numbers to the design hypotheses stated in Section 3.5.

### 4.1. Real Data

Tables 1 and 2 report the metrics for caption prediction and concept detection on real data. The Overall column is the official average over the six caption metrics.

**Run 1.** Greedy decoding yields our team's highest concept F1 score at 0.5594 together with a secondary score of 0.9465. The result confirms the hypothesis stated in Section 3.5: the adapted model emits the dominant CUI set with high confidence under the simplest deterministic decoding policy, and no further sophistication is required on the concept side. As anticipated, greedy generation was not viable for caption submission, and no caption score is therefore available for this run.

**Run 2.** Adding the deduplication post-processing of Section 3.5 to the Run 2 output yields the highest overall caption score among our submissions, at 0.3065. The largest gain over Run 2 is on ROUGE-1,

**Table 3**
Caption prediction results on synthetic data. Columns and ordering follow Table 1.

| Run | ID | Overall | BERT | ROUGE | Similarity | BLEURT | UMLS Concept F1 | Align |
|---|---|---|---|---|---|---|---|---|
| 1 | 788 | 0.3095 | 0.5247 | 0.2670 | 0.7838 | 0.3388 | 0.1264 | 0.1543 |
| 2 | 876 | 0.3031 | 0.5459 | 0.2695 | 0.7521 | 0.3211 | 0.1214 | 0.1466 |
| 3 | 884 | 0.3024 | 0.5344 | 0.2610 | 0.7784 | 0.3148 | 0.1197 | 0.1454 |
| 4 | 920 | 0.2959 | 0.5393 | 0.2639 | 0.7663 | 0.3165 | 0.1187 | 0.1220 |

which rises from 0.1862 to 0.2224. The increase reflects closer alignment between the generated captions and the concise structure of the reference captions, since semantic duplicates that inflated the n-gram count without adding clinical content have been removed. Concept F1 is identical to Run 2 at 0.5540, which confirms that the post-processing operates only on the caption text and leaves the CUI list intact. The configuration therefore validates the hypothesis stated for Run 2 in Section 3.5.

**Run 3.** Beam-2 decoding without post-processing yields an overall caption score of 0.3011. The configuration produces the highest image-caption similarity of our submissions at 0.7845 and a competitive BLEURT of 0.3404. The wider search yields more coherent captions than greedy decoding, but it does not by itself remove the surface repetition that affects the n-gram-based metrics. ROUGE-1 remains low at 0.1862. Concept F1 is essentially unchanged from Run 1 at 0.5540, which confirms that the dominant CUIs are already covered by the top hypothesis under beam search.

**Run 4.** Stochastic sampling at raised temperature degrades both subtasks, and the trade-off anticipated in Section 3.5 resolves on the unfavorable side. Concept F1 falls from 0.5540 to 0.3407, and the overall caption score drops from 0.3011 to 0.2797. The factuality components are particularly affected, with UMLS Concept F1 at 0.1080 and AlignScore at 0.1123. The flattened next-token distribution admits clinically plausible but visually unsupported CUIs, and the increased lexical diversity does not produce a compensating gain on the relevance side. The result indicates that medical captioning in this setting favors deterministic decoding.

**Run 5.** Beam-4 decoding with the n-gram repetition penalty and strict truncation produces the highest BERTScore of the campaign at 0.5669 together with a ROUGE-1 of 0.2217, close to that of Run 3. The configuration therefore reaches its design goal of dense and grammatically complete captions. The risk anticipated in Section 3.5 also materializes: the strict truncation removes valid CUIs that the model emits at the end of the sequence, and concept F1 drops to 0.3084 as a result. The configuration trades concept recall for caption fluency. The overall caption score falls to 0.2859, below Run 2, despite the gain on BERTScore, because the official average is dominated by the lower factuality components.

## 4.2. Synthetic Data

Tables 3 and 4 report the metrics for the synthetic subtasks. The same checkpoint is used as on the real subtasks, so the results characterize zero-shot transfer from real to synthetic radiology images.

**Run 1.** The beam-3 configuration with repetition penalty reaches the highest synthetic overall caption score at 0.3095 and the highest image-caption similarity at 0.7838. Both values are close to the corresponding real-data numbers obtained by Run 2 on the real side. Concept F1, however, is 0.0552, one order of magnitude below the real-data baseline of 0.5594. The result indicates an asymmetric domain gap. The language model retains its syntactic and lexical capacity under the synthetic distribution, since the caption metrics are essentially unaffected. The visual encoder, in contrast, maps synthetic textures to UMLS embeddings with substantially lower confidence, and the resulting CUI predictions are

**Table 4**
Concept detection results on synthetic data. F1 is defined as in Table 2. Runs are ordered by F1.

| Run | ID | F1 |
|---|---|---|
| 2 | 832 | 0.1048 |
| 1 | 786/789 | 0.0552 |
| 4 | 918 | 0.0542 |
| 3 | 883 | 0.0530 |

diffuse rather than concentrated on the correct concepts. The two submissions with this configuration produced identical scores, which indicates that submission-level variance is negligible relative to the configuration effects studied here.

**Run 2.** Filtering CUI predictions by a probability threshold approximately doubles concept F1 to 0.1048, which is the best synthetic concept score among our submissions. The same configuration also yields the highest ROUGE-1 of the synthetic campaign at 0.2695, at a modest cost in overall caption score (0.3031) and image-caption similarity (0.7521). The result supports the hypothesis stated for Run 2 in Section 3.5: the synthetic-domain shift inflates the rate of spurious low-confidence concepts, and removing them improves concept F1 without disturbing the high-confidence content of the captions. The configuration therefore identifies probability calibration as an effective low-cost mitigation for the synthetic concept gap.

**Run 3.** Length-constrained generation produces dense captions but reverts concept F1 to 0.0530. Restricting `max_new_tokens` to 90 prevents the model from emitting the longer-tail CUIs that are needed to recover recall on the synthetic distribution. The gain anticipated on caption density does not materialize either: the overall caption score is 0.3024, essentially equal to that of Run 2. The configuration therefore fails on both objectives, and the secondary risk anticipated in Section 3.5 dominates the intended effect.

**Run 4.** Using the later QLoRA adapter state instead of the earlier adapter state did not improve the synthetic results. The overall caption score was 0.2959 and the concept F1 was 0.0542, both below the corresponding values obtained by Run 1, despite using the strict-truncation post-processing inherited from Run 5 on the real data. Between the two outcomes anticipated in Section 3.5, the result supports the second hypothesis: additional adaptation on the real-image distribution further specialized the adapter to real visual patterns and slightly reduced transfer to synthetic images. This configuration therefore suggests that extending training on real data alone is not an effective mitigation for the synthetic-domain gap in concept detection.

Three patterns emerge from the submitted runs. First, on real data, the optimal decoding strategy differs between sub-tasks. Greedy decoding maximizes concept F1, while beam search with sentence-level deduplication maximizes the caption overall score. Second, stochastic sampling degrades both sub-tasks, which is consistent with the prior observation that medical generation favors deterministic decoding. Third, the transfer from real to synthetic data exposes an asymmetric domain gap. The overall caption score is essentially preserved between the two distributions, but concept F1 drops by almost an order of magnitude. Confidence-threshold filtering recovers a substantial fraction of the synthetic concept F1, while extended training on real data does not.

## 5. Conclusions

We presented the NUSTPB system for ImageCLEFmedical Caption 2026, which applies QLoRA adaptation to LLaVA-1.5-7B and produces modality, CUIs, and findings from a single instruction prompt. By holding

the adapted weights fixed across all submissions, our experiments isolate the effect of decoding strategy and post-processing on the official evaluation metrics. The unified prompt design supports concept detection and caption prediction with one model and one inference pass per image, and the same checkpoint serves the real and synthetic subtasks in accordance with the data-separation rule. The results identify three actionable findings. Greedy decoding is the strongest choice for concept detection on real data, while beam search combined with sentence-level deduplication is the strongest choice for caption prediction; selecting decoding strategy per subtask is therefore preferable to a single shared configuration. Stochastic decoding is consistently harmful in this domain and should be avoided. The dominant remaining bottleneck is the visual side of the synthetic-domain gap on concept detection, which is not closed by additional training on real data. Probability-based filtering of CUI predictions partially recovers performance on synthetic data and provides a low-cost mitigation, but a fuller solution likely requires synthetic-domain adaptation of the visual encoder or the projection layer, which we leave to future work.

## Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools, specifically ChatGPT and Claude, to support sentence-level improvements, including grammar and spelling correction and limited paraphrasing for clarity and conciseness. The use of these tools was restricted to drafting support and did not involve the generation of scientific content, results, or conclusions. All AI-assisted outputs were carefully reviewed, edited, and verified by the authors. The authors take full responsibility for the content and integrity of the final manuscript.

## References


[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, 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. Eryılmaz, 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. Eryılmaz, 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] B. Ionescu, H. Müller, M. Villegas, H. Arenas, G. Boato, D.-T. Dang-Nguyen, Y. D. Cid, C. Eickhoff, A. G. S. de Herrera, C. Gurrin, M. B. Islam, V. A. Kovalev, V. Liauchuk, J. Mothe, L. Piras, M. Riegler, I. Schwall, Overview of imageclef 2017: Information extraction from images, in: Conference and Labs of the Evaluation Forum, 2017. URL: https://api.semanticscholar.org/CorpusID:27831479.

[4] B. Ionescu, H. Müller, M. Villegas, A. García Seco de Herrera, C. Eickhoff, V. Andrearczyk, Y. Dicente Cid, V. Liauchuk, V. Kovalev, S. A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, D.-T. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, C. Gurrin, Overview of imageclef 2018: Challenges, datasets and evaluation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 309–334.

[5] B. Ionescu, H. Müller, R. Péteri, Y. D. Cid, V. Liauchuk, V. Kovalev, D. Klimuk, A. Tarasau, A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, M. Lux, C. Gurrin, O. Pelka, C. M. Friedrich, A. Garcìa Seco de Herrera, N. Garcia, E. Kavallieratou, C. R. del Blanco, C. Cuevas, N. Vasillopoulos, K. Karampidis, J. Chamberlain, A. Clark, A. Campello, Imageclef 2019: Multimedia retrieval in medicine, lifelogging, security and nature, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 358–386.

[6] B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, V. Datla, S. A. Hasan, D. Demner-Fushman, S. Kozlovski, V. Liauchuk, Y. D. Cid, V. Kovalev, O. Pelka, C. M. Friedrich, A. García Seco de Herrera, V.-T. Ninh, T.-K. Le, L. Zhou, L. Piras, M. Riegler, P. Halvorsen, M.-T. Tran, M. Lux, C. Gurrin, D.-T. Dang-Nguyen, J. Chamberlain, A. Clark, A. Campello, D. Fichou, R. Berari, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, Overview of the imageclef 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2020, pp. 311–341.

[7] B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. D. Cid, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the imageclef 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2021, pp. 345–370.

[8] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. B. Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the imageclef 2022: Multimedia retrieval in medical, social media and nature applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 541–564.

[9] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W.-W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. M. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, I. Coman, V. Kovalev, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of the imageclef 2023: Multimedia retrieval in medical, social media and internet applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 370–396.

[10] B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A.-G. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W.-W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of the imageclef 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 140–164.

[11] 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. B. Abacha, A. G. S. 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. Eryılmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, M. S. Hee, 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, S. Anand, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: J. Carrillo-de Albornoz, A. García

Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2026, pp. 290–314.

[12] J. Rückert, A. Ben Abacha, A. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. Pakull, et al., Overview of imageclefmedical 2024–caption prediction and concept detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740, CEUR-WS, 2024, pp. 1437–1455.

[13] H. Damm, T. M. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, et al., Overview of imageclefmedical 2025–medical concept detection and interpretable caption generation, Proceedings of the CLEF, Madrid, Spain (2025) 9–12.

[14] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2023) 34892–34916.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.

[16] 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.