Overview of ImageCLEFmedical 2026 – Medical Concept Detection and Caption Generation with Synthetic Data Extensions

Hendrik Damm[1,2,*,†], Tabea M. G. Pakull[3,1,†], Lea Reinartz[1,2], Benjamin Bracke[1,2], Bahadır Eryılmaz[4], Praveen Nath[1], Raphael Brüngel[1,2,4], Cynthia S. Schmidt[4], Henning Schäfer[1,4], Asma Ben Abacha[6], Alba G. Seco de Herrera[7], Henning Müller[8,9] and Christoph M. Friedrich[1,2]

[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany*

[2]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany*

[3]*Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany*

[4]*Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Germany*

[6]*Microsoft, Redmond, Washington, USA*

[7]*School of Computer Science, National University of Distance Education (UNED), Spain*

[8]*University of Applied Sciences Western Switzerland (HES-SO), Switzerland*

[9]*University of Geneva, Switzerland*

Abstract

The 10th Edition of the ImageCLEFmedical Caption task extends the established concept detection and caption prediction benchmark with two synthetic-data subtasks and a manually evaluated explainability task. The 2026 release is built from finalized 2025 outputs and expanded with new PubMed Central data and updated annotations. The main dataset contains 131 853 images (97 364 train, 19 240 validation, 15 249 test) and 2 646 unique UMLS Concept Unique Identifiers (CUIs). The synthetic-caption dataset contains 131 723 images (97 222 train, 19 239 validation, 15 262 test) and 2 174 unique CUIs. Both datasets were harmonized through missing-image pruning, zero-concept removal, and CUI-to-name mappings. The final test phase received 102 finished runs across four prediction leaderboards, and the explainability task received three manually evaluated submissions. Best hidden-test scores were 0.5790 F1-score for main concept detection, 0.3755 overall for main caption prediction, 0.5609 F1-score for synthetic concept detection, and 0.5840 overall for synthetic caption prediction. The best explainability submission reached an overall manual score of 3.69 on a 1–5 scale. This paper documents the 2026 task design, data preparation, synthetic-data extension, evaluation setup, organizer baselines, and final results.

Keywords

ImageCLEF, Computer Vision, Multi-Label Classification, Image Captioning, Vision-Language Models, Radiology, Synthetic Data

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

*Corresponding author.

†These authors contributed equally.

hendrik.damm@fh-dortmund.de (H. Damm); tabea.pakull@uk-essen.de (T. M. G. Pakull); lea.reinartz@fh-dortmund.de (L. Reinartz); benjamin.bracke@fh-dortmund.de (B. Bracke); bahadir.eryilmaz@uk-essen.de (B. Eryılmaz); raphael.bruengel@fh-dortmund.de (R. Brüngel); cynthia.schmidt@uk-essen.de (C. S. Schmidt); henning.schaefer@uk-essen.de (H. Schäfer); abenabacha@microsoft.com (A. Ben Abacha); alba.garcia@lsi.uned.es (A. G. Seco de Herrera); henning.mueller@hevs.ch (H. Müller); christoph.friedrich@fh-dortmund.de (C. M. Friedrich)

0000-0002-7464-4293 (H. Damm); 0009-0009-9802-7167 (T. M. G. Pakull); 0009-0000-8014-7624 (L. Reinartz); 0000-0003-4986-7142 (B. Bracke); 0009-0002-8743-4751 (B. Eryılmaz); 0000-0002-6046-4048 (R. Brüngel); 0000-0003-1994-0687 (C. S. Schmidt); 0000-0002-4123-0406 (H. Schäfer); 0000-0001-6312-9387 (A. Ben Abacha); 0000-0002-6509-5325 (A. G. Seco de Herrera); 0000-0001-6800-9878 (H. Müller); 0000-0001-7906-0038 (C. M. Friedrich)

1. Introduction

ImageCLEF[1] [1] is the image retrieval and classification lab of the Conference and Labs of the Evaluation Forum (CLEF) [2]. Within ImageCLEF, the ImageCLEFmedical Caption task has become a recurring benchmark for the automatic understanding of radiology images, in which systems must both identify the clinical concepts present in an image and describe its content in fluent natural language.

The Caption task was first introduced in 2016 and, in 2017 and 2018 [3, 4], comprised the two subtasks of concept detection and caption prediction. From 2019 [5] to 2020 [6], the focus narrowed to concept detection, that is, extracting Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) from radiology images. Since 2021 [7], both subtasks have run in parallel again, supported by gradually higher-quality, manually curated annotations and, from 2023 [8], a switch from BLEU to BERTScore [9] as the primary caption-prediction metric. In 2025 [10], the caption evaluation was broadened to a composite of six relevance and factuality measures, and a manually evaluated explainability subtask was introduced.

Building on these developments, 2026 marks the 10^{th} edition of the ImageCLEFmedical Caption task. Its motivation is unchanged: the manual creation of structured knowledge from medical images is slow, costly, and error-prone, so benchmarking systems that detect clinical concepts, compose fluent radiology captions, and justify their outputs continues to drive research toward scalable and trustworthy radiology-image understanding. As in previous editions, the development data are derived from an extended release of the Radiology Objects in COntext version 2 (ROCOv2) dataset [11], which is built from PubMed Central Open Access articles [12].

The central novelty of the 2026 edition is a synthetic-data extension. Alongside the two established subtasks on real captions and the explainability subtask, we introduce two synthetic-data subtasks in which models are evaluated on a companion dataset whose captions are generated under controlled, strictly image-grounded constraints. The goal is twofold: to study model behavior under the stylistic and lexical distribution shift between human-written and synthetically generated captions, and to provide a more uniform captioning target next to the noisier real-caption setting. A second, infrastructural change concerns reproducibility: the yearly data pipeline was rebuilt so that the final exported files of one edition serve directly as the input basis for the next. This design reduces manual handover effort and makes the release process more reproducible.

This paper presents an overview of the ImageCLEFmedical Caption 2026 task: the task design and participation (Section 2), the data creation and synthetic extension (Section 3), the evaluation methodology (Section 4), the organizer baselines (Section 5), the results (Section 6), and the conclusions (Section 7).

2. Task and Participation

The ImageCLEFmedical Caption 2026 challenge is organized into five subtasks:

1. **T1 Main Concept Detection**: predict UMLS CUIs from real radiology images.
2. **T2 Main Caption Prediction**: generate one caption per real radiology image.
3. **T3 Synthetic Concept Detection**: predict UMLS CUIs from synthetic-caption dataset images.
4. **T4 Synthetic Caption Prediction**: generate captions for synthetic-caption dataset images.
5. **T5 Explainability**: provide human-interpretable explanations for caption outputs on a small subset of test images.

Within each dataset family, concept detection and caption prediction share the same image-only test input package. Participants therefore receive one main test-image zip and one synthetic test-image zip, while submissions differ only in the required CSV format.

[1] https://www.imageclef.org/ [last accessed: 2026-05-12]

2.1. Final Participation Statistics

Table 1 summarizes the final hidden-test phase. Because the leaderboard export is account-based and one spreadsheet explicitly marks at least one alias relation between accounts, we report submitting-account counts exactly as exported by the platform.

Table 1
Final-phase participation statistics for ImageCLEFmedical Caption 2026.

Task	Accounts	Finished runs/submissions	Primary metric
T1 Main Concept Detection	12	35	F1
T2 Main Caption Prediction	13	43	Overall(F/R)
T3 Synthetic Concept Detection	3	7	F1
T4 Synthetic Caption Prediction	7	17	Overall(F/R)
T5 Explainability	3	3	Manual 1–5 score
Total	–	105	–

Beyond the leaderboard submissions, 16 teams contributed accompanying working-note papers documenting their approaches this year. Notably, 12 of the 16 participants made their source code available, suggesting increased adherence to open science practices.

3. Data Creation and Release

3.1. Main Dataset (Real Captions)

Data originate from PubMed Central Open Access articles and are processed using the yearly CLEF pipeline built on ROCOv2 [11, 12]. For 2026, the split construction follows the yearly handover logic: 2025 train and validation data are merged into the 2026 training partition, the 2025 test split becomes the 2026 validation split, and newly introduced 2026 images form the new hidden test split. All exported IDs are regenerated in the 2026 namespace. Table 2 reports the resulting split statistics after final quality control.

Table 2
Main dataset split statistics after final quality control.

Split	Rows/Images	Unique CUIs	Zero-CUI rows	One-CUI rows	≥2-CUI rows	Avg CUIs/image
Train	97 364	2 646	0	9 902	87 462	3.21
Valid	19 240	2 428	0	1 912	17 328	3.27
Test	15 249	2 307	0	1 464	13 785	3.35
Total	131 853	2 646	0	13 278	118 575	3.24

All final test concepts are contained in the training vocabulary (test concepts not in train: 0).

The modality distribution of the main dataset is reported in Table 3; computed tomography (DRCT) and X-ray (DRXR) together account for roughly 62% of all images.

Table 4 details the IRMA body-region distribution for the X-ray (DXR) subset of the main dataset.

Table 3
Main dataset modality distribution.

Modality	Images	Percent
DRCT	47 178	35.781%
DRXR	35 216	26.709%
DRMR	21 213	16.088%
DRUS	19 262	14.609%
DRAN	6 547	4.965%
DRCO	1 153	0.874%
DRPE	840	0.637%
DROC	444	0.337%

Table 4
IRMA-region distribution for DXR images in the main dataset. Percentages are calculated over images with available IRMA region labels ($N = 35\,095$).

IRMA label	Images	Percent
chest	12 272	34.968%
cranium	6 030	17.182%
lower extremity (leg)	5 037	14.352%
abdomen	3 801	10.831%
upper extremity (arm)	2 458	7.004%
pelvis	2 180	6.212%
spine	1 939	5.525%
other	1 139	3.245%
breast (mamma)	239	0.681%
Total	35 095	100.000%

3.2. Synthetic Dataset

The synthetic dataset is created from a secondary caption source in which each main-dataset image is paired with a synthetically generated caption. We keep the source-image split assignment fixed (train remains train, validation remains validation, test remains test), then assign shuffled 2026 synthetic IDs inside each split.

Concepts are extracted from synthetic captions with the same MedCAT-to-UMLS pipeline used for the main track. As in the main dataset, rows with zero final CUIs are removed before export.

3.2.1. Synthetic Caption Generation Pipeline

Synthetic captions are generated in a dedicated workflow outside the yearly CLEF packaging pipeline and then merged back into the final 2026 synthetic release. The generation workflow has four stages:

1. **NXML context extraction.** For each image, the pipeline locates the corresponding PubMed Central article source, links the figure references in the article body to the matching figure image, and extracts both the original figure-caption text and a local paragraph window around each figure mention.
2. **Prompted VLM caption generation.** The extracted context, legacy figure-caption text, and metadata (modality and IRMA region) are passed to a Qwen3-VL family model served through vLLM. Generation is executed asynchronously with resume support, retry/backoff handling, and incremental checkpoint writes for fault tolerance on long runs.
3. **Prompt-level hallucination control.** Generation is steered by a fixed system prompt and a per-image user prompt template rather than by a short free-form instruction. The prompt design enforces strict image-grounding (positive findings only, no speculative clinical claims) and

standardized modality/region wording; for chest X-ray cases, the finding vocabulary is aligned with CheXpert/CheXbert-style labels [13]. The complete system and user prompt templates are reproduced in Appendix B.

4. **Quality review and integration.** A random caption sample is exported as an HTML review pack (image + source caption + synthetic caption + PMC link) for manual auditing before final integration into the synthetic caption release.

This workflow adapts core NXML figure-caption extraction ideas from the Biomedica ETL ecosystem [14] to the ImageCLEFmedical yearly release process. The resulting split statistics after quality control are reported in Table 5.

Table 5
Synthetic dataset split statistics after final quality control.

Split	Rows/Images	Unique CUIs	Zero-CUI rows	One-CUI rows	≥2-CUI rows	Avg CUIs/image
Train	97 222	2 174	0	1 063	96 159	5.04
Valid	19 239	2 087	0	151	19 088	5.14
Test	15 262	1 977	0	137	15 125	5.08
Total/Union	131 723	2 174	0	1 351	130 372	5.06

As shown in Table 6, the modality distribution of the synthetic dataset closely mirrors that of the main dataset, since both are built from the same underlying source images.

Table 6
Synthetic dataset modality distribution.

Modality	Images	Percent
DRCT	47 169	35.809%
DRXR	35 215	26.734%
DRMR	21 210	16.102%
DRUS	19 112	14.509%
DRAN	6 547	4.970%
DRCO	1 173	0.891%
DRPE	840	0.638%
DROC	398	0.302%
DROCT (legacy label)	59	0.045%

For IRMA in the synthetic track, we provide source-linked IRMA labels inherited from the corresponding main-dataset source image.

3.3. Quality Control and Released Packages

- Main dataset rows removed during final zero-concept pruning: 45 (train 0, valid 13, test 32).
- Synthetic dataset rows removed during final zero-concept pruning: 175 (train 142, valid 14, test 19).
- Missing-image pruning in the main dataset removed 5 rows before final packaging.
- Final Codabench-style packages are generated for main and synthetic tracks, each with dev and test bundles.

CUI mapping release. To support participants, we release resolved CUI-to-name dictionaries for both datasets: one for the main dataset, covering all 2 646 concepts, and one for the synthetic dataset, covering all 2 174 concepts. All CUIs in the final datasets were successfully resolved to names via UMLS MRCONSO.

4. Evaluation Methodology

The evaluation procedure builds on the revised methodology introduced in 2025 [10]. As in previous editions, the subtasks were evaluated independently. The standard tasks (T1, T2) and the newly introduced synthetical tasks (T3, T4) share the same evaluation procedures unless noted otherwise. The source code of all evaluation scripts is publicly available.[2] The repository also provides scripts to check submission formats.

In 2026, the AI4MediaBench[3] by AIMultimediaLab[4] was used as the challenge platform.

4.1. Concept Detection (T1, T3)

For the concept detection subtasks, the balanced precision and recall trade-off were measured in terms of F1-scores. A secondary F1-score is computed using a subset of concepts that was manually curated. On the one hand, this involves the different image modalities (X-ray, Angiography, Ultrasound, CT, MRI, PET, OCT, and Combined such as PET/CT). On the other hand, if applicable, for X-ray also the anatomical code for body region examined of IRMA (cranium, chest, upper extremity, spine, abdomen, pelvis, breast, and lower extremity) was involved. There is no secondary score for the Synthetical Concept Detection subtask (T3).

Submissions follow the strict format `ID,CUIs` with semicolon-separated, unique CUIs per row and at most 100 CUIs per image. All image IDs from the test set must be present in the submission file.

The ground truth for the test set was generated based on the same reduced subset of the UMLS 2022 AB release that was used for the training data (see Section 3 for more details).

4.2. Caption Prediction (T2, T4)

For caption prediction, system outputs were assessed using a composite score, averaging across six complementary metrics to jointly capture aspects of relevance and factuality. All individual scores for each caption are summed and averaged over the number of captions, resulting in the final score. Submissions use the strict format `ID,Caption` (UTF-8, one row per test ID, CSV-safe quoting).

4.2.1. Relevance

Relevance was evaluated using four different methods. The first of these is BERTScore [9], which is a metric that computes a similarity score for each token in the generated text with each token in the reference text. It uses the pre-trained contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) [15]-based models and matches words by cosine similarity. In this work, the pre-trained model *microsoft/deberta-xlarge-mnli*[5] was used because it is the model that correlates best with human scoring according to the authors. Following best practices for caption evaluation reported by [9], we computed Recall-based BERTScore with inverse document frequency (idf) weighting, using idf scores derived from the test set to emphasize informative terms.

The second metric, ROUGE (Recall-Oriented Understudy for Gisting Evaluation [16]) score, counts the number of overlapping units such as n-grams, word sequences, and word pairs between the generated text and the reference. Specifically, the ROUGE-1 (F-measure) score was calculated, which measures the number of matching unigrams between the model-generated text and a reference.

The third relevance metric BLEURT (BiLingual Evaluation Understudy with Representations from Transformers) [17] is designed to assess the quality of natural language generation in English by leveraging a pre-trained model that has been fine-tuned to emulate human judgments about the quality

[2] https://github.com/taubsity/clef-caption-evaluation-2026 last access 2026-07-04.
[3] https://ai4media-bench.aimultimedialab.ro/ last access 2026-07-04.
[4] https://www.aimultimedialab.ro/ last access 2026-07-04.
[5] https://huggingface.co/microsoft/deberta-xlarge-mnli last access 2026-07-04.

of the generated text. The strength of BLEURT lies in its end-to-end training, which enables it to model human judgments effectively and makes it robust to domain and quality variations. For this evaluation, the BLEURT-20 model was used.

All of the above-mentioned metrics were computed using preprocessed captions that were lowercased and had punctuation stripped. Numeric values were replaced with the token "number." The captions were treated as single sentences, regardless of actual sentence boundaries. This step ensures uniformity and focuses the evaluation on linguistic content.

In addition to the text-based metrics a reference-free metric was implemented. The methodology is based on CLIPScore [18], an innovative metric that diverges from the traditional reference-based evaluations of image captions. Instead, it aligns with the human approach of evaluating caption quality without references by evaluating the alignment between text and image content. The original metric employs Contrastive Language-Image Pretraining (CLIP) [19], a cross-modal model that has been pre-trained on a massive dataset of image-caption pairs sourced from the web. For this tast evaluation the MedImageInsight [20] model was used instead. It is trained using medical images with associated text and labels from a variety of domains, including X-ray, CT, MRI, OCT, and ultrasound. The model is used to compute similarity scores between images and text.

4.2.2. Factuality

To assess the factuality of the generated captions, two complementary metrics were employed. The UMLS Concept F1-score evaluates the overlap of medical entities between the generated and reference captions. Specifically, medical concepts were extracted using MedCAT [21], with a focus on semantic types relevant to clinical accuracy as also defined for the MEDCON [22] metric whereas MEDCON relies on QuickUMLS [23] for concept extraction from both texts. This is followed by calculation of the F1-score to quantify concept-level agreement. The other factuality metric, AlignScore [24], employs a deep learning approach based on RoBERTa [25] to measure factual consistency. It involves the decomposition of extensive texts into more manageable segments and aligning the claims in the generated caption with the supporting evidence in the reference caption, thereby producing an average alignment score across all claims.

4.3. Explainability (T5)

The explainability task asks participants to provide explanations for captions on a small subset of test images. Participants were encouraged to be creative and were not constrained to a fixed output format.

A radiologist manually evaluated each submission on eight test images using nine 1–5 Likert-scale criteria, with 5 being the best score.

The captions were rated in terms of readability, accuracy, level of detail, and focus. The readability scale ranks whether the predicted captions are readable and coherently formulated. The accuracy evaluates whether the predicted captions match ground-truth captions or are clinically plausible from the clinician's perspective. The level of detail is used to assess whether the captions merely describe visual findings or also interpret underlying clinical concepts. The focus validates the appropriateness of the scope of the caption and thus penalizes short captions that lack essential observations as well as excessively long captions that are not focused on the essentials.

The visualization was assessed based on consistency, comprehensiveness, and focus. The consistency measures whether the visualization is consistent with the predicted caption, e.g. whether the correct number of annotations matches the described concepts. The comprehensiveness scale assesses whether the visualizations cover all relevant concepts. The focus validates the appropriateness of the scope of the visualisation.

Each image was rated individually and an average score across categories was reported. In addition, the radiologist rated the appropriateness of the overall methodology and provided a clinician's overall favorite rating. The final score was calculated as the unweighted mean over all nine criteria.

5. Organizer Baselines

We report organizer baselines on the public validation leaderboards so they are comparable and reproducible independently of the hidden test phase.

5.1. Concept Baseline (Main and Synthetic)

We train an image-only multi-label classifier using tf_efficientnetv2_s with ImageNet pretraining, BCEWithLogitsLoss, and class-wise pos_weight.

Training setup: Image size 300, batch size 48, learning rate 2×10^{-4}, weight decay 10^{-4}, 10 epochs. During inference, probabilities are thresholded and constrained to at most 100 CUIs per image. Threshold tuning is performed on validation over $\{0.60, 0.65, \ldots, 0.95\}$.

Table 7
Organizer concept baseline official validation scores (AI4MediaBench).

Dataset	Date (UTC)	Submission ID	F1	F1 (secondary, manual)	Task version
Main	2026-02-16 11:29	53	0.45	0.89	Concept Detection Valid (2026) v2
Synthetic	2026-02-16 12:25	58	0.38	0.00	Concept Detection Valid (2026) Synthetic v2

Internal threshold tuning selected threshold 0.95 for both main and synthetic runs.

5.2. Caption Baseline (Main and Synthetic)

We fine-tune Qwen3-VL-4B-Instruct with Unsloth (4-bit loading and PEFT adapters). The baseline is trained and queried with a single short instruction, *"You are a medical imaging expert. Describe this radiology image in one concise sentence."*, which is deliberately kept distinct from the elaborate prompt templates used to generate the synthetic-caption dataset (Appendix B).

Training setup: 1.5 epochs, batch size 1, gradient accumulation 8, learning rate 2×10^{-4}, max length 1024, max new tokens 96.

Table 8
Organizer caption baseline official validation scores (AI4MediaBench).

Dataset	Date (UTC)	ID	Overall (F/R)	Relevancy	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
Main	2026-02-19 12:16	102	0.3427	0.5344	0.1511	0.6086	0.2710	0.9324	0.3257	0.1722	0.1299
Synthetic	2026-02-20 10:09	109	0.5514	0.6935	0.4094	0.7364	0.6029	0.9428	0.4921	0.4872	0.3315

Training runtime and setup details are reported in the run reports; the table above reflects official public validation results.

6. Results

This section reports the final hidden-test leaderboards. Main-task captioning and concept detection attracted the largest number of runs, while the synthetic concept track remained the smallest leaderboard. Throughout the result tables we preserve the platform-exported account names.

6.1. T1 Main Concept Detection

The main concept-detection leaderboard was highly competitive. The leading F1-score of 0.5790 was achieved by `dsgt_caption` (run 647), with `archimedes` close behind at 0.5781. The best-per-account top six all scored at least 0.5721 F1-score. The top secondary score among the best-per-account entries was 0.9657, again achieved by run 647.

Table 9
T1 main concept detection: best final result per submitting account.

Account	Run ID	F1	Secondary F1
dsgt_caption	647	0.5790	0.9657
archimedes	983	0.5781	0.9591
UIT_Worker	608	0.5725	0.9533
basharderar	869	0.5725	0.9419
AUEB/DMST	1003	0.5721	0.9503
nhanbui	657	0.5713	0.9553
DeepLens	959	0.5648	0.9336
dingglebellw	757	0.5606	0.9428
NUSTPB	765	0.5594	0.9465
krishnatewari	661	0.5234	0.8403
UIT_HighDefinition	755	0.2739	0.9349

dsgt_caption [26] Team dsgt_caption ranked 1st with an F1-score of 0.5790 and a secondary F1-score of 0.9657. This was achieved by implementing a late-fusion ensemble of ConvNeXt-V2-Base [27], BiomedCLIP ViT-B/16 [28], and DenseNet-169 [29] models using a regularized "Honest Threshold Tuning" procedure that aimed to avoid overfitting on rare concepts during validation. The pipeline behind this procedure comprised a deterministic split, long-tail fallback, and threshold blending to yield a concept-individual threshold instead of applying a single global one. As for the models, the team employed ConvNeXt-V2-Base for down-weighting of well-classified negative samples, BiomedCLIP ViT-B/16 for strong medical visual priors from PubMed Central pre-training, and DenseNet-169 for architectural diversity. In addition, they tested a training-free KNN retrieval pipeline over frozen BiomedCLIP embeddings that performed nearly on a par with their winning approach.

archimedes [30] The joint AUEB NLP Group and NKUA Archimedes Unit team built on their past work by developing a highly complex ensemble of multiple ImageNet-initialized CNNs (EfficientNet [31] variants, ConvNeXt-Tiny [32], DenseNet-121 [29], and ResNet-50 [33]) combined with FFNN classifiers. They employed intricate ensemble strategies including techniques like soft-voting raw probability scores, Dual-L thresholding, and Bidirectional Conformal Rescue that marginally improved results. The team secured second place overall with a primary F1-score of 0.5781 and a secondary F1-score of 0.9591.

UIT_Worker [34] The UIT_Worker team proposed a simpler but carefully optimized convolutional neural network pipeline over complex multi-stage architectures or ensembles. The approach utilized a single DenseNet-121 [29] initialized with standard ImageNet weights, combined with a full fine-tuning strategy (training + validation datasets), and Asymmetric Loss [35] to downweight easy negative samples to better handle extreme long-tail label dataset imbalance. The team secured third place with a primary F1-score of 0.5725 and a secondary F1-score of 0.9533.

basharderar [36] Team basharedar introduced a frequency-aware multi-head strategy to address the broad and imbalances multi-label setting of the given dataset. This was employed by decomposing the concept space into frequency-based groups, balancing the impact of frequent concepts against rare ones. The implementation is based on a SwinTransformer [37] backbone, trained with Asymmetric Loss [35] and further stabilization techniques for improved robustness. Their approach achieved an F1-score of 0.5725 equal to that of team UIT_Worker, and a secondary F1-score of 0.9419, ranking 4th.

AUEB/DMST [38] The AUEB/DMST team employed a weighted soft-voting ensemble of CNN and Vision Transformer backbones, including ConvNeXt, Swin-Small, DINOv2, and EfficientNetV2, for multi-label UMLS concept detection, achieving a primary F1-score of 0.5721 and a secondary F1-score of 0.9503 on the standard concept detection task.

DeepLens [39] The contribution of the DeepLens team displays a continuation of their experiments from their prior participation [40] in ImageCLEFmedical Caption 2025. They re-employed their EfficientNet [31] B0 and B1 models and fine-tuned their DenseNet [29] model on the current dataset. In addition, they evaluated an ensemble of their models. Of these approaches, their EfficientNet B1 model performed best, achieving 7th rank with an F1-score of 0.5648 and a secondary F1-score of 0.9336.

NUSTPB [41] The NUSTPB team adapted the LLaVA-1.5-7B vision-language model using QLoRA [42] for joint medical concept detection and caption generation, where deterministic greedy decoding achieved the best concept detection performance with an F1-score of 0.5594 and a secondary F1-score of 0.9465.

krishnatewari [43] The proposed approach is based on a retrieval framework using DenseNet121 image embeddings. Visually similar images are retrieved using cosine similarity, and their concept labels are aggregated using a weighted k-nearest neighbour strategy. The team achieved an F1-score of 0.5234 and a secondary F1-score of 0.8403.

UIT_HighDefinition [44] The UIT_HighDefinition team trained a DenseNet121 [29] multi-label classifier over UMLS concepts, initialized with ImageNet weights and optimized with AdamW and an asymmetric multi-label loss [35] to address class imbalance. Sigmoid probabilities were thresholded at 0.60, achieving an F1-score of 0.2739.

6.2. T2 Main Caption Prediction

For the main caption task, `AIstatLab` led the hidden-test leaderboard with an overall score of 0.3755. The next best best-per-account submissions were `UFPE-Essex-UNED` at 0.3603 and `dsgt_caption` at 0.3571. Across the best-per-account rows, relevance scores were substantially higher than factuality scores, indicating that factual precision remained the main bottleneck.

Table 10
T2 main caption prediction: best final result per submitting account.

Account	Run ID	Overall	Relevancy	Factuality
AIstatLab	937	0.3755	0.5786	0.1724
UFPE-Essex-UNED	977	0.3603	0.5527	0.1679
dsgt_caption	990	0.3571	0.5365	0.1777
AUEB/DMST	985	0.3516	0.5288	0.1743
csmorgan	936	0.3458	0.5377	0.1539
gfreitas	840	0.3343	0.5201	0.1484
UIT_HighDefinition	980	0.3193	0.4999	0.1388
NUSTPB	774	0.3065	0.4705	0.1426
UMUTeam	806	0.2990	0.4798	0.1181
DS-MED Lab	964	0.2814	0.4526	0.1102
krishnatewari	674	0.2685	0.4346	0.1025
UACH-VisionLab	1004	0.2673	0.4275	0.1072
nguyenthinh	780	0.2527	0.4373	0.0680

AIstatLab [45] The AIstatLab team combined supervised fine-tuning with a post-SFT Group Reward-Decoupled Policy Optimization (GDPO) stage. A multi-alignment reward was organized around three axes implemented with four components: BERTScore Recall and ROUGE-1 for caption-to-caption alignment, BiomedCLIP similarity for caption-to-image alignment, and a frozen Biomed-BERT caption-to-CUI extractor with a Soft F1-score reward. Heterogeneous rewards were balanced through reward-decoupled advantage normalization and a manually designed piecewise weight schedule. The submission used a Qwen-MedVL-2B model with MedGemma-4B as an additional candidate source, beam-search decoding, optional RadGraph-XL-guided reranking,

and a conservative rule-based fallback, ranking 1st with an overall score of 0.3755.

UFPE-Essex-UNED [46] The UFPE-Essex-UNED team developed a Retrieval-Augmented Generation (RAG) framework for image captioning. They used MedSigLIP [47] to embed images and retrieve relevant exemplar captions. These captions were then inserted into the Llama 4 [48] prompt.

dsgt_caption [26] Team dsgt_caption ranked 3rd with an overall score of 0.3571, a relevancy score of 0.5365, and the highest factuality score of 0.1777. This was yielded by their best performing approach on caption prediction, based on a fine-tuned Gemma-3 27B [49] model. They further conducted experiments with a fine-tuned BLIP pipeline with custom Vizwins merging, and a zero-shot MedGemma-4B [47] with PubMed-style prompting.

AUEB/DMST [38] The AUEB/DMST team developed a dual-encoder-decoder captioning architecture combining BiomedCLIP [28], Swin-Small [50], a Q-Former bridge, and a LoRA-adapted FLAN-T5 decoder, achieving an overall caption score of 0.3516 on the standard caption prediction task.

csmorgan [51] The team fine-tuned LLaVA-OneVision-Qwen2 at 0.5B and 7B parameter scales using the parameter-efficient adaptation methods QLoRA and QDoRA to create four systems. Their best model, 7B-QDoRA, performed the best across all evaluation metrics and was ranked 5th overall.

gfreitas [52] The team evaluated three prompting strategies - one-shot, random few-shot and Retrieval-Augmented Generation (RAG)-based few-shot - across the MedGemma and Qwen3-VL models. For the RAG approach, they used MedSigLIP embeddings to provide contextual examples to guide caption generation. Qwen3-RAG achieved the team's best overall score of 0.3343.

UIT_HighDefinition [44] The UIT_HighDefinition team used an approach of concept-enriched prompting with Qwen3-VL-8B-Instruct [53] to generate captions. DenseNet121-base [29] predicted the concepts, which were included in the prompt for the generator with the instruction to include them only if they were visually supported. The caption is then post-processed to normalize modality expressions, remove generic prefixes, control length, and validate formatting of the submission.

NUSTPB [41] The NUSTPB team employed a QLoRA-adapted LLaVA 1.5-7B framework for unified caption generation and concept prediction, where beam-search decoding with sentence-level deduplication achieved the best overall caption score of 0.3065.

UMUTeam [54] The UMUTeam built a lightweight vision-language model from a frozen CLIP-ViT-Large-Patch14 encoder, a Q-Former-style projection module, and a Gemma-3-1B-it language model adapted with LoRA. Only the projector and the LoRA adapters were trained (about 1.29% of all parameters), while the visual encoder remained frozen, and the model was trained exclusively on the synthetic caption subset. Conditioned on the projected visual embeddings and a fixed "Describe the image:" prompt, the system reached an overall score of 0.2990 with a relevance of 0.4798 and a factuality of 0.1181, ranking 9th of 13 submissions.

DS-MED Lab [55] The DS-MED Lab team conditioned a fine-tuned BLIP-large decoder on explicit semantic priors. A BiomedCLIP-based multimodal clustering module (PCA and K-Means) assigned each image to one of three clinical macro-domains, and a dual-branch EfficientNet-B4 classifier with Class-Specific Residual Attention and Asymmetric Loss [35] performed multi-label CUI detection. Predicted concepts and routing tokens were injected through a cluster-aware prompt with Scheduled Teacher Forcing Dropout, ranking 10th.

krishnatewari [43] The team developed an encoder-decoder architecture consisting of an InceptionV3 image encoder and a two-layer LSTM decoder. They also explored the use of an alternative Qwen2.5-0.5B decoder experimentally. Captions were generated using a beam search approach incorporating repetition penalties to enhance fluency and minimise repetitive outputs. The system achieved an overall score of 0.2685.

UACH-VisionLab [56] The framework proposed by the UACH-VisionLab combines a controlled two-

stage caption preprocessing pipeline with a multimodal image captioning architecture. Captions are first cleaned and paraphrased using LLMs to reduce non-visual information and reporting inconsistencies, creating an enhanced training dataset. The captioning model then leverages a frozen PubMedCLIP image encoder, a Transformer-based mapping network, and a finetuned GPT-2 decoder to generate clinically meaningful medical image descriptions, which partly deviate from the goldstandard.

6.3. T3 Synthetic Concept Detection

The synthetic concept leaderboard showed a clear leader. `AUEB/DMST` reached 0.5609 F1-score, ahead of `krishnatewari` at 0.4741.

Table 11
T3 synthetic concept detection: best final result per submitting account.

Account	Run ID	F1
AUEB/DMST	951	0.5609
krishnatewari	684	0.4741
NUSTPB	832	0.1048

AUEB/DMST [38] The AUEB/DMST team employed a weighted soft-voting ensemble of CNN and Vision Transformer backbones, including ConvNeXt, Swin-Small, DINOv2, and EfficientNetV2, for multi-label UMLS concept detection, achieving the best overall performance on the synthetic concept detection task with an F1-score of 0.5609.

krishnatewari [43] The team used their retrieval-based framework built on DenseNet121 image embeddings. This framework retrieves visually similar images using cosine similarity and aggregates their concept labels through a weighted k-nearest neighbour strategy. The system achieved an F1-score of 0.4741.

NUSTPB [41] The NUSTPB team evaluated the transferability of their QLoRA-adapted LLaVA-1.5-7B model to synthetic radiology data, where probability-threshold filtering on predicted UMLS concepts achieved a synthetic concept detection F1-score of 0.1048.

6.4. T4 Synthetic Caption Prediction

The synthetic caption track produced the strongest caption-generation scores of the whole benchmark. `AIstatLab` achieved the top hidden-test score with 0.5840 overall, followed by `AUEB/DMST` at 0.5136. Compared with the main caption task, both relevance and factuality scores are substantially higher on the synthetic track.

Table 12
T4 synthetic caption prediction: best final result per submitting account.

Account	Run ID	Overall	Relevancy	Factuality
AIstatLab	999	0.5840	0.7183	0.4496
AUEB/DMST	1005	0.5136	0.6542	0.3731
UMUTeam	803	0.4632	0.6148	0.3116
UFPE-Essex-UNED	829	0.4625	0.6372	0.2878
krishnatewari	760	0.4341	0.5758	0.2924
madhab	801	0.3861	0.5200	0.2522
NUSTPB	788	0.3095	0.4786	0.1404

AIstatLab [45] The AIstatLab team applied the same supervised fine-tuning plus post-SFT GDPO pipeline to the synthetic track, replacing the piecewise reward schedule with dynamic reward-advantage agreement statistics. Using Qwen-MedVL-2B with dynamic-scheduled GDPO and the rule-based fallback, the team ranked 1st with an overall score of 0.5840.

AUEB/DMST [38] The AUEB/DMST team applied their dual-encoder-decoder captioning framework based on BiomedCLIP, Swin-Small, Q-Former, and LoRA-adapted FLAN-T5 to the synthetic caption prediction task, achieving an overall score of 0.5136 and ranking second overall.

UMUTeam [54] The UMUTeam reused its synthetic-data-only vision-language model (frozen CLIP encoder, Q-Former-style projector, and LoRA-adapted Gemma-3-1B-it) on the synthetic caption track, where the training and evaluation distributions coincide. The team's primary interest was methodological, examining whether synthetic captions can serve as a validation proxy for the real captioning setting. The model performed markedly better here than on the main task, reaching an overall score of 0.4632 with a relevance of 0.6148 and a factuality of 0.3116, ranking 3rd of 7 submissions. The team attributed the gap between the two tracks to a distribution shift between synthetic and real captions, most visible in ROUGE and concept-level (MedCAT) factuality.

UFPE-Essex-UNED [46] The UFPE-Essex-UNED team developed a Retrieval-Augmented Generation (RAG) framework for image captioning. They used CLIP ViT-B/32 [19] to embed images and retrieve relevant exemplar captions. These captions were then inserted into the Llama 4 [48] prompt.

krishnatewari [43] The team used their encoder-decoder architecture consisting of an InceptionV3 image encoder and a two-layer LSTM decoder. They also explored the use of an alternative Qwen2.5-0.5B decoder. Captions were generated using a beam search approach incorporating repetition penalties to enhance fluency and minimise repetitive content. The system achieved an overall score of 0.4341.

NUSTPB [41] The NUSTPB team applied the same QLoRA-adapted LLaVA-1.5-7B framework to synthetic caption prediction in a zero-shot setting, achieving an overall caption score of 0.3095 using beam-search-based decoding strategies.

6.5. T5 Explainability

Three teams submitted explainability outputs for manual review. The highest overall manual score was achieved by Mahmudul Hoque from Morgan State University with a LLaVA-OneVision-Qwen2 system using QLoRA/QDoRA (3.69). Krishna Tewari from the Indian Institute of Technology submitted a phrase-grounding and counterfactual XAI approach (2.85), and `gfreitas` submitted a MedGemma intrinsic-XAI and interactive-HTML approach (2.83). The radiologist's notes highlighted Team 2's polished and clinically accurate captions, while also noting that its final graphical explanation component was less explicit than in some other submissions.

Table 13
T5 explainability task: manually evaluated final submissions. Overall score is the mean over nine 1–5 radiologist-rated criteria.

Team	Participant	Affiliation	Approach	Overall score
Team 2	Mahmudul Hoque	Morgan State University	LLaVA-OneVision-Qwen2 + QLoRA/QDoRA	3.69
Team 1	Krishna Tewari	Indian Institute of Technology	Phrase grounding + counterfactual XAI	2.85
Team 3	Gabriel Freitas	–	MedGemma intrinsic XAI + interactive HTML	2.83

csmorgan (Mahmudul Hoque) [51] The team proposed an explainability framework based on model-intrinsic attention that extracts attention directly from the captioning model's decoder, thereby explaining the model's reasoning process. The relevant image regions are visualised

using heatmaps and bounding boxes. The framework achieved first place overall, winning six out of nine evaluation criteria.

krishnatewari (Krishna Tewari) [43] The proposed framework unifies clinical phrase extraction, CLIPSeg-based phrase grounding, evidence aggregation, counterfactual analysis and Qwen2-VL-based explanation generation. The final output is a comprehensive clinical XAI panel integrating visual and textual explanations. This methodology achieved second place overall. However, it should be noted that the explainability approach focuses solely on the generated captions and does not involve the black-box model itself. This means that it does not enhance radiologists' trust in the model's predictions.

gfreitas (Gabriel Freitas) [52] The proposed method is an explainability framework based on intrinsic embedding that links image patches and caption chunks using the cosine similarity of their multimodal embeddings. The chunks are mapped onto a spatial grid to create a heatmap. The team then produced an interactive HTML visualisation in which hovering over a caption segment dynamically highlights the corresponding image regions.

7. Conclusion and Discussion

For the 10th edition of ImageCLEFmedical Caption, we provide a cleaned and reproducible yearly pipeline, a large updated main dataset, and a new synthetic companion dataset with dedicated concept and caption subtasks. The released data are internally coherent across captions, concepts, attributions, IDs, and images after strict quality filtering.

Compared with prior editions, 2026 introduces three practical improvements for participants: explicit synthetic-track benchmarking, released CUI-to-name mapping files for all final concepts, and a manually evaluated explainability task. On the hidden test phase, the best main-task results reached 0.5790 F1-score for concept detection and 0.3755 overall for caption prediction. On the synthetic track, the best concept score reached 0.5609 F1-score and the best caption score reached 0.5840 overall, suggesting that synthetic caption generation produced a benchmark that remains non-trivial for concept prediction while being substantially easier for caption generation. In the explainability task, the best manually evaluated submission reached 3.69 on a 1–5 scale.

The synthetic extension behaved differently from the established tracks, and the contrast is informative. As a newly introduced setting, the synthetic subtasks attracted markedly fewer participants: the synthetic concept and caption leaderboards drew 3 and 7 submitting accounts (7 and 17 finished runs), against 12 and 13 accounts (35 and 43 runs) on the corresponding main tracks. The synthetic leaderboards also show steeper score drop-offs, so the signal beyond the leading systems is thinner than on the main tracks.

For concept detection, the synthetic task proved almost as hard as the real one: the best synthetic F1-score of 0.5609 trails the best main F1-score of 0.5790 only slightly, even though synthetic images carry more concepts per image on average (5.06 versus 3.24) drawn from a smaller vocabulary (2 174 versus 2 646 unique CUIs). Detecting clinical concepts thus remains a genuinely difficult visual problem irrespective of whether the supervising captions are human-written or synthetic. Caption prediction shows the opposite picture: the best synthetic overall score of 0.5840 far exceeds the best main score of 0.3755, with both relevance and factuality higher on the synthetic track. The organizer baselines reproduce this pattern (synthetic caption baseline 0.5514 versus 0.3427 on the main track), indicating that the effect originates in the target itself rather than in any particular participant system.

We attribute this gap to the regularity of the synthetic captions. Generated under strict image-grounding constraints with standardized modality and region wording and a controlled finding vocabulary, they are stylistically uniform, free of non-visual context, and close to the text current vision-language models naturally produce. Real captions instead carry author-specific phrasing, non-visual clinical context, and reporting variability that depress both relevance and factuality. High scores on the synthetic caption track should therefore be read as evidence of target regularity rather than of easier clinical content; as

several participants observed, the distribution shift between synthetic and real captions—most visible in ROUGE and concept-level (MedCAT) factuality—means that synthetic performance is not a drop-in proxy for the real-caption setting. Taken together, the two families form a useful pair: concept detection transfers well between them, while the synthetic caption track isolates linguistic-style effects and offers a cleaner, more controlled companion benchmark for future editions.

Acknowledgments

The work of Benjamin Bracke, Lea Reinartz and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Henning Schäfer, Tabea M. G. Pakull, Hendrik Damm, Praveen Nath, and Bahadır Eryılmaz was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed). This work was partly supported by the projects GRESEL-UNED (PID2023-151280OB-C22) funded by MICIU/AEI/ AEI 501100011033 and ANNOTATE (PID2024-156022OB-C31) funded by MICIU/AEI/10.13039/ 501100011033 and the European Social Fund Plus (ESF+).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to support grammar and spelling checks. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication content.

References

[1] H. Müller, J. Kalpathy-Cramer, A. García Seco de Herrera, Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks, Springer International Publishing, Cham, 2019, pp. 231–250. URL: https://doi.org/10.1007/978-3-030-22948-1_10. doi:10.1007/978-3-030-22948-1_10.

[2] N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF, Springer, Cham, 2019. URL: https://link.springer.com/book/10.1007/978-3-030-22948-1. doi:10.1007/978-3-030-22948-1.

[3] C. Eickhoff, I. Schwall, A. García Seco de Herrera, H. Müller, Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017. URL: http://ceur-ws.org/Vol-1866/invited_paper_7.pdf.

[4] A. García Seco de Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018. URL: http://ceur-ws.org/Vol-2125/invited_paper_4.pdf.

[5] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2019 concept detection task, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_245.pdf.

[6] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding, in: CLEF2020 Working Notes, volume 1166 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2020.

[7] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, pp. 1101–1112.

[8] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – caption prediction and concept detection, in: CLEF2023 Working Notes, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2023, pp. 1328–1346.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

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

[11] 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.

[12] R. J. Roberts, PubMed Central: The GenBank of the published literature, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 381–382. doi:10.1073/pnas.98.2.381.

[13] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, M. P. Lungren, Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, arXiv preprint arXiv:2004.09167 (2020). URL: https://arxiv.org/abs/2004.09167.

[14] A. Lozano, M. W. Sun, J. Burgess, L. Chen, J. J. Nirschl, J. Gu, I. Lopez, J. Aklilu, A. W. Katzer, C. Chiu, A. Rau, X. Wang, Y. Zhang, A. S. Song, R. Tibshirani, S. Yeung-Levy, Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 19724 – 19735. doi:10.1109/cvpr52734.2025.01837.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[16] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[17] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. doi:10.18653/v1/2020.acl-main.704.

[18] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. doi:10.18653/v1/2021.emnlp-main.595.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[20] N. C. F. Codella, Y. Jin, S. Jain, Y. Gu, H. H. Lee, A. Ben Abacha, A. Santamaria-Pang, W. Guyman, N. Sangani, S. Zhang, H. Poon, S. Hyland, S. Bannur, J. Alvarez-Valle, X. Li, J. Garrett, A. McMillan, G. Rajguru, M. Maddi, N. Vijayrania, R. Bhimai, N. Mecklenburg, R. Jain, D. Holstein, N. Gaur,

V. Aski, J.-N. Hwang, T. Lin, I. Tarapov, M. Lungren, M. Wei, MedImageInsight: An open-source embedding model for general domain medical imaging, 2024. URL: https://arxiv.org/abs/2410.06542. arXiv:2410.06542.

[21] 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.

[22] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation, Scientific Data 10 (2023) 586. doi:10.1038/s41597-023-02487-3.

[23] L. Soldaini, N. Goharian, QuickUMLS: A fast, unsupervised approach for medical concept extraction, in: Medical Information Search Workshop (MEDIR) at SIGIR, Pisa, Italy, 2016.

[24] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. doi:10.18653/v1/2023.acl-long.634.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[26] B. Wang, Y. Zhang, R. Mehta, DS@GT ARC at ImageCLEFmedical 2026: Architectural Diversity for Concept Detection and Foundation-Model Scaling for Caption Prediction in Medical Image Analysis, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[27] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2023, p. 16133–16142. URL: http://dx.doi.org/10.1109/cvpr52729.2023.01548. doi:10.1109/cvpr52729.2023.01548.

[28] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, H. Poon, A multimodal biomedical foundation model trained from fifteen million image–text pairs, NEJM AI 2 (2024). doi:10.1056/AIoa2400640.

[29] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[30] S. Stavrou, D. Boutzounis, A. Chatzipapadopoulou, F. Charalampakos, K. V. Dalakleidi, J. Pavlopoulos, Y. Panagakis, Uncertainty-Aware Biomedical Image Concept Detection via CNN Ensembles, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[31] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning (ICML 2019), 2019, pp. 6105–6114.

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[34] T. Q. To, D. C. Pham, N. H. Bui, T. B. Nguyen-Tat, UIT_Worker at ImageCLEFmedical 2026: DenseNet-121 Full-Fit Medical Concept Detection with Asymmetric Loss, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[35] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: 2021 IEEE/CVF International Conference on Computer Vision

(ICCV), 2021, pp. 82 – 91. doi:10.1109/iccv48922.2021.00015.
[36] B. Alsaleh, R. Sonbol, Z. Baghdadi, Frequency-Aware Multi-Head Learning for Long-Tailed Medical Concept Detection in ImageCLEFmed Caption 2026, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2021, p. 9992–10002. URL: http://dx.doi.org/10.1109/iccv48922.2021.00986. doi:10.1109/iccv48922.2021.00986.
[38] G. Karandreas, P. Louridas, AUEB/DMST at ImageCLEFmedical caption 2026: Concept detection and caption prediction with cnn-vit ensembles and dual-encoder architectures, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[39] B. K. Nejad, A. H. S. Rudsari, S. Eetemadi, FDeepLens at ImageCLEFmedical 2026: A Continuation of CNN-Based Medical Concept Detection, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[40] A. H. S. Rudsari, B. K. Nejad, M. Hajihosseini, S. Eetemadi, Detecting concepts for medical images: Contributions of the DeepLens team at IUST to ImageCLEFmedical caption 2025, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[41] M. C. Tatomir, D. C. Stanciu, L. D. Ştefan, NUSTPB at ImageCLEFmed caption 2026: Medical image captioning with QLoRA-adapted LLaVA-1.5, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[42] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in Neural Information Processing Systems 36 (2023).
[43] K. Tewari, P. D. N. V. Sohana, S. Pal, Seeing, Describing, and Explaining Medical Image Understanding via Retrieval and Multimodal Learning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[44] N. T. Duy, N. T. Q. Han, N. M. Triet, T. B. Nguyen-Tat, UIT_HighDefinition at ImageCLEFmedical caption 2026: Soft CUI-guided Qwen3-VL captioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[45] H. J. Kim, H. Shin, M. Kim, C. Lim, GDPO-based multi-alignment reward optimization for clinically grounded medical image captioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[46] H. M. Pereira, P. D. M. ao, V. de A. Xavier, T. I. Ren, A. G. Seco de Herrera, V. Abolghasemi, UFPE-Essex-UNED at ImageCLEFcaption 2026: Training-free multimodal RAG for medical image captioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[47] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden, L. Yang, Medgemma technical report, 2026. URL: https://arxiv.org/abs/2507.05201. arXiv:2507.05201.
[48] Meta AI, The llama 4 herd: The beginning of a new era of natively multimodal intelligence, 2025. URL: https://ai.meta.com/blog/llama-4-multimodal-intelligence/, accessed: 2025-06-05.
[49] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian,

M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[51] M. Hoque, R. N. Chowdhury, O. O. E. Peter, O. Islam, R. Hoque, M. Rahman, Model-Intrinsic Attention as Explanation for Radiology Image Captioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[52] G. Freitas, M. Bosco, R. Lotufo, J. Pereira, D. Carmo, Retrieval Augmented Generation and Patch-Level Explainability applied to Medical Image Captioning with Small Visual Language Models, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[53] 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, 2025. URL: https://arxiv.org/abs/2511.21631. arXiv:2511.21631.

[54] R. Pan, P. J. Vivancos-Vicente, J. Gómez-Navalón, T. Bernal-Beltrán, J. A. García-Díaz, UMUTeam at ImageCLEF 2026: Exploring synthetic data for medical image captioning with a Q-Former-based vision-language model, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[55] T.-T. Nguyen-Chi, B. Q. P. Ho, T. B. Nguyen-Tat, uit-01100010en at ImageCLEFmedical caption 2026: Clinical-aware routing for hallucination suppression and pathology-focused medical image captioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[56] S. Rascon-Cervantes, G. Ramirez-Alonso, D. R. Lopez-Flores, A. P. López-Monroy, Improving image–text alignment in medical image captioning through textual preprocessing and multimodal conditioning, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

A. Full Results

A.1. T1 Main Concept Detection

Table 14
Full final leaderboard for T1 main concept detection.

Account	Run ID	F1	Secondary F1
dsgt_caption	647	0.5790	0.9657
archimedes	983	0.5781	0.9591
dsgt_caption	903	0.5780	0.9599
archimedes	818	0.5780	0.9590
archimedes	984	0.5776	0.9614
dsgt_caption	1002	0.5774	0.9558
archimedes	931	0.5771	0.9574
dsgt_caption	880	0.5768	0.9556
archimedes	987	0.5764	0.9601
y_zhang	606	0.5763	0.9634
UIT_Worker	608	0.5725	0.9533
basharderar	869	0.5725	0.9419
UIT_Worker	972	0.5725	0.9529
AUEB/DMST	1003	0.5721	0.9503
AUEB/DMST	817	0.5719	0.9514
nhanbui	657	0.5713	0.9553
UIT_Worker	971	0.5712	0.9546
AUEB/DMST	954	0.5697	0.9426
UIT_Worker	663	0.5669	0.9413
DeepLens	959	0.5648	0.9336
DeepLens	928	0.5632	0.9323
DeepLens	925	0.5632	0.9323
dingglebellw	757	0.5606	0.9428
dingglebellw	756	0.5602	0.9421
NUSTPB	765	0.5594	0.9465
DeepLens	927	0.5542	0.9276
NUSTPB	773	0.5540	0.9364
NUSTPB	776	0.5540	0.9364
nhanbui	648	0.5419	0.8750
UIT_Worker	662	0.5314	0.8901
krishnatewari	661	0.5234	0.8403
dsgt_caption	1001	0.4893	0.8863
NUSTPB	854	0.3407	0.5371
NUSTPB	955	0.3084	0.4900
UIT_HighDefinition	755	0.2739	0.9349

A.2. T2 Main Caption Prediction

Table 15: Full final leaderboard for T2 main caption prediction (core metrics).

Account	Run ID	Overall	Relevancy	Factuality	BERT	ROUGE
AIstatLab	937	0.3755	0.5786	0.1724	0.6250	0.3234
AIstatLab	993	0.3754	0.5749	0.1759	0.6258	0.3216
AIstatLab	816	0.3753	0.5784	0.1722	0.6247	0.3231
AIstatLab	922	0.3747	0.5794	0.1700	0.6266	0.3250
AIstatLab	945	0.3730	0.5748	0.1712	0.6243	0.3194
UFPE-Essex-UNED	977	0.3603	0.5527	0.1679	0.6017	0.2725

Account	Run ID	Overall	Relevancy	Factuality	BERT	ROUGE
dsgt_caption	990	0.3571	0.5365	0.1777	0.6042	0.2809
dsgt_caption	841	0.3564	0.5351	0.1776	0.6095	0.2736
UFPE-Essex-UNED	989	0.3554	0.5468	0.1641	0.5976	0.2668
AUEB/DMST	985	0.3516	0.5288	0.1743	0.6130	0.2728
AUEB/DMST	879	0.3492	0.5304	0.1679	0.6146	0.2735
csmorgan	936	0.3458	0.5377	0.1539	0.6075	0.2725
dsgt_caption	901	0.3446	0.5247	0.1644	0.6049	0.2743
AUEB/DMST	875	0.3443	0.5210	0.1676	0.6097	0.2695
UFPE-Essex-UNED	825	0.3433	0.5408	0.1459	0.5903	0.2527
gfreitas	840	0.3343	0.5201	0.1484	0.6013	0.2226
csmorgan	910	0.3343	0.5257	0.1429	0.6022	0.2689
gfreitas	826	0.3339	0.5147	0.1530	0.5887	0.2346
csmorgan	913	0.3325	0.5247	0.1403	0.5997	0.2597
dsgt_caption	649	0.3324	0.5134	0.1515	0.5960	0.2588
csmorgan	912	0.3317	0.5230	0.1403	0.5978	0.2580
gfreitas	827	0.3282	0.4957	0.1608	0.5956	0.2210
UIT_HighDefinition	980	0.3193	0.4999	0.1388	0.5977	0.2243
dsgt_caption	878	0.3186	0.5120	0.1253	0.5908	0.2472
UIT_HighDefinition	979	0.3169	0.4821	0.1516	0.5888	0.2372
AUEB/DMST	831	0.3124	0.4574	0.1674	0.5616	0.2226
NUSTPB	774	0.3065	0.4705	0.1426	0.5465	0.2224
UIT_HighDefinition	863	0.3011	0.4880	0.1143	0.5865	0.2267
NUSTPB	775	0.3011	0.4528	0.1495	0.4999	0.1862
UMUTeam	806	0.2990	0.4798	0.1181	0.5670	0.2230
UIT_HighDefinition	837	0.2934	0.4764	0.1103	0.5902	0.2289
UIT_HighDefinition	842	0.2918	0.4761	0.1075	0.5830	0.2265
NUSTPB	956	0.2859	0.4609	0.1110	0.5669	0.2217
DS-MED Lab	964	0.2814	0.4526	0.1102	0.5690	0.2126
NUSTPB	855	0.2797	0.4492	0.1101	0.5309	0.1876
krishnatewari	674	0.2685	0.4346	0.1025	0.5398	0.2059
UACH-VisionLab	1004	0.2673	0.4275	0.1072	0.5742	0.1818
UACH-VisionLab	621	0.2626	0.4152	0.1100	0.5786	0.1727
UACH-VisionLab	860	0.2615	0.4233	0.0996	0.4911	0.1613
UACH-VisionLab	861	0.2561	0.4267	0.0855	0.4965	0.1380
nguyenthinh	780	0.2527	0.4373	0.0680	0.5720	0.2207
nguyenthinh	976	0.2515	0.4230	0.0800	0.5507	0.1684
UACH-VisionLab	718	0.2397	0.3996	0.0798	0.4914	0.1500

A.3. T3 Synthetic Concept Detection

Table 16
Full final leaderboard for T3 synthetic concept detection.

Account	Run ID	F1	Secondary F1
AUEB/DMST	951	0.5609	0.0000
krishnatewari	684	0.4741	0.0000
NUSTPB	832	0.1048	0.0000
NUSTPB	789	0.0552	0.0000
NUSTPB	786	0.0552	0.0000
NUSTPB	918	0.0542	0.0000
NUSTPB	883	0.0530	0.0000

A.4. T4 Synthetic Caption Prediction

Table 17: Full final leaderboard for T4 synthetic caption prediction (core metrics).

Account	Run ID	Overall	Relevancy	Factuality	BERT	ROUGE
AIstatLab	999	0.5840	0.7183	0.4496	0.7580	0.6353
AIstatLab	828	0.5839	0.7189	0.4488	0.7576	0.6343
AIstatLab	995	0.5839	0.7182	0.4496	0.7577	0.6352
AIstatLab	807	0.5838	0.7189	0.4488	0.7576	0.6342
AIstatLab	809	0.5736	0.7106	0.4367	0.7486	0.6227
AUEB/DMST	1005	0.5136	0.6542	0.3731	0.7195	0.5555
AUEB/DMST	904	0.5083	0.6512	0.3653	0.7172	0.5531
AUEB/DMST	952	0.5081	0.6480	0.3683	0.7096	0.5504
UMUTeam	803	0.4632	0.6148	0.3116	0.6930	0.5127
UFPE-Essex-UNED	829	0.4625	0.6372	0.2878	0.6621	0.4795
krishnatewari	760	0.4341	0.5758	0.2924	0.6687	0.4758
madhab	801	0.3861	0.5200	0.2522	0.6033	0.3889
NUSTPB	788	0.3095	0.4786	0.1404	0.5247	0.2670
NUSTPB	785	0.3095	0.4786	0.1404	0.5247	0.2670
NUSTPB	876	0.3031	0.4722	0.1340	0.5459	0.2695
NUSTPB	884	0.3024	0.4722	0.1326	0.5344	0.2610
NUSTPB	920	0.2959	0.4715	0.1203	0.5393	0.2639

A.5. T5 Explainability

Table 18
Detailed manual averages for the T5 explainability task. All criteria use a 1–5 scale.

Team	Readability	Accuracy	Detail	Caption focus	Consistency	Coverage	Vis. focus	Methodology	Favorite	Overall
Team 1	2.625	2.125	2.750	2.625	2.625	3.125	2.750	4.000	3.000	2.847
Team 2	4.625	3.625	3.375	4.250	3.125	2.875	3.375	3.000	5.000	3.694
Team 3	4.500	3.125	3.625	4.125	1.000	2.125	1.000	3.000	3.000	2.833

B. Synthetic Prompt Templates

This appendix lists the prompt templates used for synthetic caption generation.

B.1. System Prompt

You are a medical imaging captioning assistant generating GOLD-STANDARD figure captions for a radiology captioning benchmark.

Benchmark constraint (critical)

- The participant will ONLY see the image.
- Therefore, the caption must contain ONLY information directly verifiable from the image pixels or burned-in text, markers, or overlays.
- Provided paper context and any existing figure caption are ONLY hints to help notice what is present; they are NOT sources of facts.

Hallucination control (critical)

- If you are not confident a detail is visible, omit it.
- Do NOT add measurements, diagnoses, demographics, or clinical history unless explicitly written on the image.
- Do NOT explain causes, etiology, or treatment; do NOT add probabilities, model outputs, or performance

claims.

Positive-only rule (critical; enforce strictly)

- Write ONLY what *is* present or visible.
- NEVER write non-findings, absence, exclusions, or rule-outs.
- NEVER use negation or exclusion language, including: `no`, `not`, `without`, `absent`, `negative for`, `rule out`, `free of`, `unremarkable`, `normal`, `clear`, `no evidence of`, `no signs of`.
- If context or an existing caption mentions a finding that is not visibly verifiable, do not negate it; simply omit it.

Caption quality target

- Write a scientific, paper-grade caption: concise, precise, image-grounded, anatomically specific.
- Prefer radiologic visual descriptors (e.g. *opacity*, *fracture line*, *effusion*) over clinical diagnoses (e.g. *pneumonia*), unless the diagnosis is literally written on the image.

Mandatory inclusions

1. Always mention the imaging type (e.g. X-ray radiograph, CT, MRI, ultrasound, angiography/fluoroscopy, PET/nuclear medicine, OCT, color optical image, or other as appropriate).
2. Always mention the IRMA body region (exactly one of: abdomen; breast (mamma); chest; cranium; lower extremity (leg); pelvis; spine; upper extremity (arm); other).

View / plane / laterality rules

- Mention projection, plane, or view (AP/PA/lateral; axial/coronal/sagittal; longitudinal/transverse) ONLY if (a) explicitly labeled in the image, or (b) unambiguous from visual presentation (high confidence).
- Mention laterality (left/right) ONLY if an L/R marker is visible or laterality is explicitly labeled.

Annotations / overlays

- If arrows, circles, boxes, labels, heatmaps, or other markings are present, mention them and the image-visible structure or finding they point to, ONLY if that indicated target is itself visible.

Editing an existing caption

- If an existing caption is provided and is image-grounded and accurate, keep it with minimal edits for style and constraints.
- If it contains non-visual, incorrect, or speculative content, rewrite completely.

Marker normalization (use exact templates; include only if visible; place at end)

- `Side marker 'R' (right) is visible.`
- `Side marker 'L' (left) is visible.`
- `Side markers 'L' (left) and 'R' (right) are visible.`
- `View label 'AP' is visible.`
- `View label 'PA' is visible.`
- `View label 'Lateral' is visible.`
- `A scale bar is present.`

Output format (strict)

- Output ONLY the final caption text.
- 1–2 sentences preferred (up to 3 if multi-panel and necessary).
- No bullet points, headings, or reasoning.

B.2. User Prompt Template

Inputs

- Paper context (hint-only; do not import non-visual facts): `{context}`
- Existing figure caption (may be empty; use only if image-grounded): `{caption_nxml}`
- Metadata (reliable; guidance only, do not output the codes verbatim):
  - Modality code: `{Modality}` (DRXR, DRMR, DRCT, DRAN, DRUS, DRPE, DRCO, DROCT, OTHER)
  - IRMA body region: `{IRMA}` (abdomen, breast (mamma), chest, cranium, lower extremity (leg), pelvis, spine, upper extremity (arm), other)

Task

Generate one paper-grade caption for the image. Every statement must be verifiable from the image pixels or burned-in text only. Report only what is present or visible; omit anything uncertain.

Modality → human-readable imaging type (use in caption)

DRXR → X-ray radiograph
DRCT → CT
DRMR → MRI
DRUS → Ultrasound
DRAN → Angiography / fluoroscopy
DRPE → PET / nuclear medicine image
DROCT → OCT (optical coherence tomography)
DRCO → Color optical image (name more specifically only if obvious)
OTHER → Other medical imaging modality (describe what it looks like)

CheXpert-style chest finding terms

Use ONLY when the modality is DRXR, the IRMA region is *chest*, and the finding is visibly present. Do not output `No Finding`; use only positive visible findings: Enlarged cardiomediastinum, Cardiomegaly, Lung lesion, Lung opacity, Edema, Consolidation, Atelectasis, Pneumothorax, Pleural effusion, Pleural other, Fracture, Support devices.

Style

- Start with imaging type + IRMA body region + view/plane if visible.
- Then salient visible findings with location, then devices or overlays.
- If multi-panel, describe panels by position (left/right/top/bottom).
- Finish with any visible marker sentences using the exact system templates.

Output

Return ONLY the final caption text.