UMUTeam at ImageCLEF 2026: Exploring Synthetic Data for Medical Image Captioning with a Q-Former-Based Vision-Language Model

UMUTeam at CLEF 2026

Ronghao Pan[1], Pedro José Vivancos-Vicente[2], Jorge Gómez-Navalón[1], Tomás Bernal-Beltrán[1] and José Antonio García-Díaz[1,*]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

[2]*VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain*

Abstract

This paper presents the participation of UMUTeam in the ImageCLEFmedical Caption 2026 task. Our work focuses on the caption prediction scenario and explores the use of synthetic caption data for training and validating medical image captioning systems. Rather than aiming only at maximizing the official ranking, the main objective of this participation was to analyze whether a model trained exclusively on synthetic captions can provide useful evidence about its expected behavior on real medical image-caption data. We propose a lightweight vision-language model based on a frozen CLIP-based visual encoder, a Q-Former-style projection module, and a Gemma-3-1B-it adapted with LoRA. The visual encoder extracts image tokens from the input medical images, while the Q-Former-style module projects these visual representations into the language model embedding space. The language model then generates captions conditioned on the projected visual embeddings and a simple textual prompt. This design allows the system to adapt to the captioning task while keeping most of the pretrained parameters frozen. The model was trained using only the synthetic caption subset. Internal validation was performed with automatic relevance-oriented metrics, including ROUGE-1 and BERTScore F1. Exact textual similarity was also computed for analysis, although its near-zero values indicate that it provides limited information for checkpoint selection in this generative setting. In the official evaluation, our approach ranked 3rd out of 7 submissions in the synthetic Caption Prediction task, obtaining an overall score of 0.4632. In the standard Caption Prediction task, the system ranked 9th out of 13 submissions, with an overall score of 0.2990. These results show that the proposed Q-Former-based VLM learns the synthetic caption distribution more effectively than it generalizes to the real captioning setting. The observed gap suggests that synthetic data can be useful for controlled model development and preliminary validation, but it should not be considered a complete substitute for validation on real medical captions.

Keywords

Vision-Language Models, Large Language Models, Natural Language Processing, Medical Image Captioning,

1. Introduction

This paper presents the participation of UMUTeam in ImageCLEF 2026, focusing exclusively on the ImageCLEFmedical Caption 2026 Caption Prediction subtask. ImageCLEFmedical Caption is a benchmark devoted to the automatic generation of textual descriptions for medical images. In this task, participating systems are required to produce captions that accurately describe the visual content of medical images while preserving the semantic and clinical relevance of the information contained in the original caption. This is a challenging problem because medical image captioning requires not only

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

*Corresponding author

ronghao.pan@um.es (R. Pan); pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); jorge.gomezn@um.es (J. Gómez-Navalón); tomas.bernalb@um.es (T. Bernal-Beltrán); joseantonio.garcia8@um.es (J. A. García-Díaz)

https://www.vocali.net/ (P. J. Vivancos-Vicente)

0009-0008-7317-7145 (R. Pan); 0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0002-3651-2660 (J. A. García-Díaz)

recognizing visual patterns, anatomical regions, imaging modalities, or pathological findings, but also expressing them in a concise and meaningful natural-language description.

In the 2026 edition, the task includes different evaluation settings, including data associated with synthetic caption generation. Our participation was limited to the caption prediction task and focused on an experimental setting based on synthetic image-caption data. Instead of designing a system whose only goal was to achieve the highest possible official score, our work was motivated by the following methodological question: to what extent can synthetic caption data support the development and preliminary validation of models intended for real medical image-caption data?

This question is relevant because annotated medical datasets are often difficult to obtain. Real medical images require expert annotation, careful anonymization, and strict ethical and privacy considerations. As a result, the availability of large-scale, high-quality image-caption datasets is limited. Synthetic data has emerged as a possible alternative to reduce this limitation, since it can provide larger volumes of annotated examples and facilitate experimentation. However, the usefulness of synthetic data depends on whether it preserves enough visual and semantic characteristics of real data. For this reason, it is important to analyze not only whether models can learn from synthetic captions, but also how their performance changes when they are evaluated in a real captioning scenario.

Our approach addresses this issue by training and validating a caption generation model using the synthetic data provided for the task. The objective was not to optimize the system only for maximum performance, but to study whether the synthetic validation subset can act as a meaningful approximation of the real captioning scenario. In other words, our participation explores the potential of the synthetic caption dataset as a proxy validation environment for real medical image captioning.

To address the proposed objective, the system follows a vision-language architecture. The visual information is extracted using a pretrained CLIP-based [1] vision encoder, which is kept frozen during training. The image representations produced by this encoder are then passed through a Q-Former-style projection module. This module transforms the visual tokens into a representation space compatible with the language model. For text generation, we use a Gemma-3-1B-it [2] causal language model adapted with LoRA, allowing efficient fine-tuning while reducing the number of trainable parameters.

A central aspect of this participation is that the synthetic data was not treated simply as a way to increase the amount of training material. Instead, it was used as the core element of the experimental design. By training and validating the model only on synthetic caption data, we aimed to examine whether this type of dataset can support the development and preliminary validation of medical captioning systems before applying them to real data. This is particularly important for scenarios where access to real annotated medical images is restricted or where synthetic data may be used as an intermediate step in model development.

In summary, this working note reports the participation of UMUTeam, which focused participation in the caption prediction task, based on a synthetic-data-driven vision-language model. The work emphasizes the validation of synthetic data as a proxy for real data, using a CLIP visual encoder, a Q-Former projection module, and a LoRA-adapted Gemma language model. The proposed methodology offers an initial step toward understanding the usefulness and limitations of synthetic medical captions for training and validating medical image captioning systems.

These working notes are structured as follows. Section 2 provides a summary of key aspects related to the task setup and background. Section 3 describes our approach in detail, outlining the architecture and components of our system. The results of our experiments are presented and discussed in Section 4. Finally, Section 5 concludes the paper with a summary of our findings and directions for future work.

2. Related Works

Medical image captioning is a multimodal task that combines computer vision and natural language generation with the objective of producing textual descriptions from medical images [3, 4]. Unlike general-domain image captioning, where the caption often describes objects, actions, and scenes in everyday language, medical image captioning requires a more specialized interpretation of visual

content [5]. The generated text should reflect relevant anatomical structures, imaging modalities, pathological findings, spatial relations, and other clinically meaningful elements [6]. This makes the task particularly challenging, since small visual differences may correspond to important semantic differences in the final caption.

The ImageCLEFmedical Caption task has become one of the main benchmarks for evaluating systems in this area. Previous editions of the task have focused on two closely related problems: medical concept detection and caption prediction. The 2024 [7], and 2025 [8] overview describes the task as aiming to extract UMLS concept annotations and/or generate captions from medical image data, with predictions compared against original image captions. It also reports the use of ROCOv2 [9] images in previous ImageCLEFmedical Caption editions. In the 2026 edition, the task continues this line of work by providing curated images from the medical literature, their captions, and associated UMLS terms, while also including a more diverse dataset to encourage more complex approaches.

Datasets such as ROCO [10] and ROCOv2 [9] have played an important role in the development of medical image captioning systems. The original ROCO dataset was introduced as a multimodal collection of radiology images designed to study the relationship between visual elements and semantic information in radiology. ROCOv2 extends this resource with radiological images, captions, and medical concepts extracted from the PMC Open Access subset. It contains 79,789 images and has been used in ImageCLEFmedical Caption tasks for concept detection and caption prediction. These datasets are especially relevant because they provide paired image-text data, which is essential for training image annotation and caption generation models in the medical domain.

Early approaches to medical image captioning often relied on encoder-decoder architectures, where a convolutional neural network extracted visual features and a recurrent or transformer-based language model generated the caption [4, 5]. More recent approaches have shifted toward large-scale vision-language models, which use pretrained visual encoders and pretrained language models to improve generalization [11, 12]. This trend is motivated by the success of transfer learning: instead of training all components from scratch, systems can reuse visual and textual representations learned from large external datasets and adapt them to the medical domain.

A particularly influential direction is the use of frozen or partially frozen vision-language architectures. BLIP-2 [13] introduced an efficient vision-language pretraining strategy that combines frozen image encoders and frozen large language models through a lightweight Querying Transformer, commonly referred to as Q-Former [13]. The Q-Former acts as a bridge between visual and textual modalities by extracting informative visual representations and projecting them into a form that can be used by the language model. This idea is closely related to our approach, where a frozen CLIP-based visual encoder is connected to a language model through a Q-Former-style projection module. In our case, this design allows the model to learn how to condition caption generation on image features without updating the entire visual encoder.

Parameter-efficient fine-tuning methods are also increasingly common in multimodal and medical language generation tasks. Large language models contain a very large number of parameters, making full fine-tuning computationally expensive. LoRA [14] addresses this issue by freezing the original model weights and injecting trainable low-rank matrices into selected layers of the transformer. This substantially reduces the number of trainable parameters while maintaining strong adaptation capability. In the context of medical image captioning, this is useful because it allows the language generation component to be adapted to the captioning task with lower computational cost and memory requirements. Our system follows this direction by applying LoRA to the Gemma-based [15] causal language model, while training only the lightweight adaptation components and the visual projection module.

Another important line of related work concerns the use of synthetic data in medical artificial intelligence. Synthetic data has been proposed as a way to address several limitations of real medical datasets, including data scarcity, privacy restrictions, annotation cost, and class imbalance. A review of synthetic data generation methods in healthcare highlights its potential to overcome data scarcity and privacy concerns, while also supporting the development of AI systems with larger and more diverse training resources [6]. In medical imaging specifically, synthetic data can enable early-stage model

development and testing without requiring direct access to sensitive patient information. However, synthetic data is not automatically equivalent to real data. Its usefulness depends on how well it preserves the statistical, visual, and semantic properties of the real target domain.

This issue is central to our participation. While many works use synthetic data mainly as an augmentation resource to improve model performance, our work focuses on a different question: whether the synthetic caption dataset can be used as a validation proxy for the real captioning task. In other words, we are interested in the representativeness of synthetic data, not only in its ability to increase the amount of available training material. This distinction is important because a model may perform well on synthetic validation data while failing to generalize to real medical captions if the synthetic data contains artifacts, simplified language, or distributions that differ from real medical literature.

Overall, previous work shows three tendencies that motivate our approach: first, the consolidation of ImageCLEFmedical Caption as a benchmark for medical image captioning; second, the adoption of pretrained vision-language architectures with lightweight modality-bridging modules such as Q-Former; and third, the growing interest in synthetic medical data as a resource for developing and validating AI systems.

Our work is positioned at the intersection of these directions. Rather than proposing only a performance-oriented captioning system, we use a CLIP visual encoder, a Q-Former-style projector, and a LoRA-adapted language model to study whether synthetic caption data can provide meaningful validation signals for the real medical captioning scenario.

3. Methodology

This section describes the experimental pipeline used for the ImageCLEFmedical Caption 2026 task. The system was developed as a vision-language captioning model trained on the synthetic caption subset. The implementation includes data preprocessing, image-text representation, a modular vision-language architecture, parameter-efficient fine-tuning, caption generation, and automatic validation. An overview of the proposed approach is shown in Figure 1.

3.1. Dataset and Experimental Setting

The experiments were conducted using the synthetic caption data associated with the task. The dataset consisted of image-caption pairs stored in a CSV file, where each sample included an image identifier and its corresponding textual caption. Images were loaded from a predefined directory, and each identifier was used to construct the path to the corresponding image file. Samples without valid identifiers or captions were removed during preprocessing.

The dataset was divided into training and validation subsets using the naming convention of the image identifiers. Samples whose identifiers contained the string `_train_` were assigned to the training subset, while those containing `_valid_` were assigned to the validation subset. This split allowed the model to be trained and evaluated on separate partitions of the synthetic dataset.

3.2. Input Representation

Each training sample consisted of one medical image and one reference caption. Images were converted to RGB format and processed with the image preprocessing pipeline associated with the visual encoder. The resulting pixel values were used as the visual input to the model.

For the textual input, a short instruction prompt was used:

```
Describe the image:
```

During training, the reference caption was appended after the prompt and tokenized using the language model tokenizer. The prompt tokens were masked in the label sequence using the value `-100`, so that the loss was computed only over the caption tokens and the end-of-sequence token.

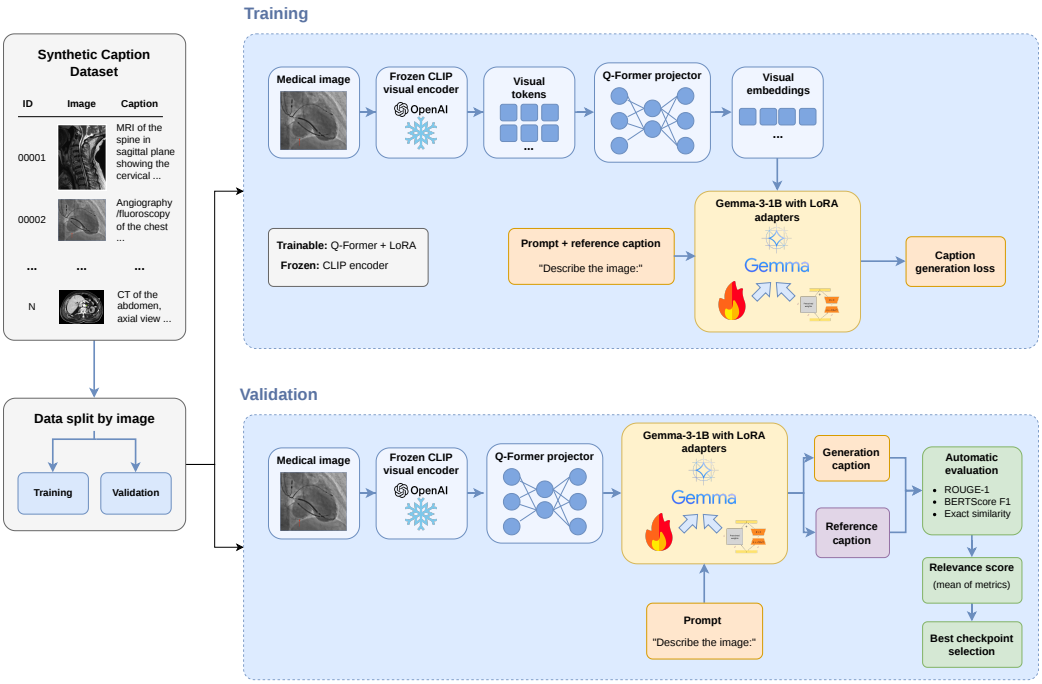

Figure 1: Overview of the proposed approach. The synthetic caption dataset is divided into training and validation subsets according to the image identifiers. During training, medical images are processed by a frozen CLIP visual encoder, projected through a Q-Former-style module, and used to condition a Gemma causal language model adapted with LoRA. During validation, generated captions are compared with reference captions using ROUGE-1, BERTScore F1, and exact textual similarity to compute the relevance score used for checkpoint selection.

The maximum text length was set to 512 tokens. Captions exceeding the available length were truncated, while shorter sequences were padded. If the tokenizer did not define a padding token, the end-of-sequence token was used as the padding token.

3.3. Model Architecture

The proposed system follows a modular vision-language architecture composed of three main components: a visual encoder, a Q-Former-style projection module, and a causal language model adapted with LoRA.

The visual encoder was based on CLIP-ViT-Large-Patch14. It was used to extract visual token representations from the input images and was kept frozen during training. This reduced the number of trainable parameters and allowed the optimization process to focus on the connection between visual representations and textual generation.

The second component was a Q-Former-style projector. This module received the visual tokens produced by the frozen vision encoder and transformed them into visual embeddings compatible with the input space of the language model. It used a fixed number of learnable query tokens, which interacted with the image tokens through self-attention and cross-attention layers. The output of the projector was then mapped to the hidden dimension of the language model.

The language generation component was based on a Gemma-3-1B-it model. Instead of full fine-tuning, LoRA adapters were applied to selected transformer layers, including attention projections and feed-forward projections. This allowed the model to adapt to the captioning task while updating only a reduced number of parameters.

The complete architecture can be summarized as follows:

1. The input image is processed by the frozen CLIP visual encoder.

2. The extracted visual tokens are passed to the Q-Former-style projector.
3. The projector maps the visual information into the language model embedding space.
4. The projected visual embeddings are concatenated with the prompt embeddings.
5. The Gemma-3-1B-it model generates the caption conditioned on the visual and textual inputs.

3.4. Training Procedure

The model was trained to generate the target caption conditioned on the image and the prompt. During the forward pass, image tokens were extracted by the frozen visual encoder and projected into the language model embedding space by the Q-Former-style module. These visual embeddings were concatenated with the textual embeddings corresponding to the prompt and the reference caption.

The attention mask was extended to include the visual tokens. Similarly, the label tensor was extended by assigning `-100` to all visual positions, ensuring that the loss was calculated only over the textual caption sequence. The model was optimized using the standard causal language modeling loss.

The Q-Former projector used 32 learnable query tokens with an internal dimension of 768, 4 transformer blocks, and 12 attention heads. The language model was adapted using LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. LoRA was applied to the attention projection layers `q_proj`, `k_proj`, `v_proj`, and `o_proj`, as well as to the MLP projection layers `gate_proj`, `up_proj`, and `down_proj`. During validation, captions were generated greedily with a maximum of 50 newly generated tokens.

Training was performed using the AdamW optimizer with weight decay. A cosine learning rate schedule with warmup was applied, and gradient clipping was used to stabilize optimization. The model was trained for five epochs with a batch size of eight, a learning rate of `1e-4`, and a warmup ratio of `0.03`.

Only trainable parameters were passed to the optimizer. These included the Q-Former-style projection module and the LoRA parameters inserted into the language model. The CLIP visual encoder remained frozen throughout training.

3.5. Evaluation Metrics

The generated captions were evaluated against the reference captions using automatic relevance-oriented metrics. Three metrics were computed:

- **ROUGE-1**, which measures unigram overlap between the generated caption and the reference caption.
- **BERTScore F1**, which estimates semantic similarity using contextual embeddings.
- **Exact textual similarity**, which checks whether the generated caption and the reference caption are identical after lowercasing.

During validation, captions were generated using the same visual-textual conditioning strategy used during training. For each image, the visual encoder extracted image tokens, the Q-Former-style module projected them into the language model space, and the prompt embeddings were concatenated after the visual embeddings. Caption generation was performed using deterministic decoding by disabling sampling. The number of newly generated tokens was limited during evaluation. After decoding, the prompt was removed from the generated output when present, leaving only the predicted caption.

The final validation score, referred to as relevance in the implementation, was computed as the mean of ROUGE-1, BERTScore F1, and exact textual similarity. This score was used to monitor the model across epochs and select the best checkpoint.

After each epoch, the model was evaluated on the validation subset. When the relevance score improved, the Q-Former weights, LoRA adapter, tokenizer, configuration, and validation metrics were saved as the best checkpoint. In addition, epoch-level checkpoints and final model artifacts were stored to preserve intermediate and final training states.

Table 1
Internal validation results across training epochs.

Epoch	Train Loss	Relevance	ROUGE-1	BERTScore
0	1.1894	0.4201	0.4947	0.7655
1	0.9866	0.4287	0.5137	0.7723
2	0.8788	0.4304	0.5176	0.7732
3	0.7783	0.4334	0.5236	0.7763
4	0.7061	0.4332	0.5233	0.7760

Table 2
Official results for UMUTeam in the Caption Prediction task.

Rank	Owner	Overall	Relevance	Factuality	BERT	ROUGE	Similarity	BLEURT	Align
9	UMUTeam	0.2990	0.4798	0.1181	0.5670	0.2230	0.8147	0.3146	0.1315

4. Results

This section reports the internal validation results obtained during training and the official results of our submission to the ImageCLEFmedical Caption 2026 task.

4.1. Training and Internal Validation

The model was trained for five epochs using the synthetic caption training subset. The training set contained 97,222 samples, while the validation subset contained 19,239 samples. The model included 1,012,931,712 total parameters, of which 13,045,760 were trainable, corresponding to 1.2879% of the total parameters. This confirms that the proposed setup follows a parameter-efficient adaptation strategy, where only the Q-Former-style projector and the LoRA adapters are updated during training.

Table 1 shows the evolution of the training loss and the internal validation metrics across epochs. The training loss decreased consistently from 1.1894 in the first epoch to 0.7061 in the final epoch. This indicates that the model was able to learn the synthetic caption distribution progressively during training.

The best internal checkpoint was obtained at epoch 3, with a relevance score of 0.4334, a ROUGE-1 score of 0.5236, and a BERTScore of 0.7763. The exact textual similarity was 0.0002, showing that the model rarely reproduced the reference captions exactly. This is expected in a generative captioning setting, where multiple valid textual formulations may describe the same image. Exact textual similarity was retained only as an auxiliary diagnostic metric; however, given its near-zero value, future versions should rely on more informative semantic and factuality-oriented metrics for checkpoint selection.

Although the training loss continued to decrease until epoch 4, the validation relevance score did not improve after epoch 3. This suggests that the model reached a performance plateau on the synthetic validation subset. Therefore, the epoch 3 checkpoint was selected as the best internal model.

4.2. Official Evaluation Results

Table 2 reports the official results for the standard Caption Prediction task. In this setting, UMUTeam ranked 9th out of 13 participating submissions. The system obtained an overall score of 0.2990, with a relevance score of 0.4798 and a factuality score of 0.1181.

The official standard caption results show that the model achieved moderate relevance but lower factuality. The relevance score indicates that the generated captions preserved part of the semantic content expected by the evaluation metrics. However, the lower factuality score suggests that the model had difficulty producing captions that were fully consistent with the clinical or concept-level information required by the real captioning setting.

Table 3 reports the official results for the synthetic Caption Prediction task. In this setting, UMUTeam ranked 3rd out of 7 participating submissions. The system obtained an overall score of 0.4632, with a

Table 3
Official results for UMUTeam in the synthetic Caption Prediction task.

Rank	Owner	Overall	Relevance	Factuality	BERT	ROUGE	Similarity	BLEURT	Align
3	UMUTeam	0.4632	0.6148	0.3116	0.6930	0.5127	0.8159	0.4376	0.2133

Table 4
Comparison between UMUTeam official results on standard and synthetic Caption Prediction.

Metric	Standard	Synthetic
Rank	9/13	3/7
Overall	0.2990	0.4632
Relevance	0.4798	0.6148
Factuality	0.1181	0.3116
BERT	0.5670	0.6930
ROUGE	0.2230	0.5127
Similarity	0.8147	0.8159
BLEURT	0.3146	0.4376
MedCAT	0.1048	0.4099
Align	0.1315	0.2133

relevance score of 0.6148 and a factuality score of 0.3116.

The system performed substantially better on the synthetic caption task than on the standard caption task. The overall score increased from 0.2990 in the standard task to 0.4632 in the synthetic task. Relevance also increased from 0.4798 to 0.6148, while factuality increased from 0.1181 to 0.3116. This difference suggests that the model was better aligned with the synthetic data distribution, which is consistent with the fact that the model was trained exclusively on synthetic captions.

4.3. Comparison Between Synthetic and Standard Caption Prediction

Table 4 summarizes the difference between the official standard and synthetic caption prediction results for UMUTeam. The comparison shows a clear performance gap between the synthetic and standard evaluation settings. The strongest differences are observed in ROUGE and MedCAT. ROUGE increased from 0.2230 in the standard task to 0.5127 in the synthetic task, suggesting a much higher lexical overlap between generated and reference captions in the synthetic setting. MedCAT also increased from 0.1048 to 0.4099, indicating that the generated synthetic captions were more consistent with the expected medical concept annotations.

The similarity score remained almost unchanged between the two settings, with 0.8147 in the standard task and 0.8159 in the synthetic task. However, the remaining metrics show that the synthetic evaluation was considerably more favorable to the proposed model. This result supports the hypothesis that the model learned patterns that transfer well within the synthetic distribution, but only partially generalize to the real caption prediction scenario.

4.4. Discussion

The results indicate that the proposed model was effective in learning from the synthetic caption data. Internally, the validation relevance score improved from 0.4201 to 0.4334 during training, and the best checkpoint was selected at epoch 3. Officially, the model achieved its strongest result in the synthetic Caption Prediction task, where UMUTeam ranked 3rd. This confirms that the model was well adapted to the synthetic caption distribution.

However, the lower performance in the standard Caption Prediction task reveals a clear distribution gap between synthetic and real data. While the synthetic results show higher relevance and factuality scores, the standard results show that the same approach does not transfer completely to real captions. In particular, the factuality score decreased from 0.3116 in the synthetic task to 0.1181 in the standard

task. This suggests that training only on synthetic captions may be insufficient for capturing the clinical variability and factual precision required by real medical image captions.

These results are aligned with the main objective of our participation. Rather than optimizing exclusively for the highest official score, our experiment aimed to analyze the extent to which synthetic caption data can support preliminary validation for real caption prediction. The ranking difference between the synthetic and standard tasks suggests that synthetic validation provides useful information about model behavior within the synthetic domain, but it should not be considered a complete substitute for validation on real medical data.

These findings should therefore be interpreted with caution. The experiment does not fully establish that synthetic validation scores reliably predict real-data performance. Instead, it shows that synthetic validation is useful for monitoring model behavior within the synthetic domain, while the gap between synthetic and standard evaluation highlights the limitations of relying exclusively on synthetic data. A stronger test of synthetic validation as a proxy would require additional experiments, such as correlating synthetic-validation scores with real-data scores across multiple checkpoints or training configurations.

5. Conclusions and further work

This paper presented the participation of UMUTeam in the ImageCLEFmedical Caption 2026 task, focusing on the caption prediction setting with particular attention to the use of synthetic caption data. The proposed system followed a modular vision-language architecture composed of a frozen CLIP visual encoder, a Q-Former-style projection module, and a Gemma causal language model adapted with LoRA. The model was trained using only the synthetic caption subset, and validation was performed through automatic relevance-oriented metrics, including ROUGE-1, BERTScore F1, and exact textual similarity.

The experimental results show that the model was able to learn meaningful patterns from the synthetic caption data. During internal validation, the relevance score improved across the first training epochs, with the best checkpoint obtained at epoch 3. In the official evaluation, the system achieved stronger results in the synthetic Caption Prediction task than in the standard Caption Prediction task. Specifically, UMUTeam ranked 3rd out of 7 submissions in the synthetic setting, with an overall score of 0.4632, while it ranked 9th out of 13 submissions in the standard caption prediction setting, with an overall score of 0.2990.

This difference between the synthetic and standard results is one of the main findings of our participation. The model obtained higher relevance, factuality, ROUGE, BERT, BLEURT, MedCAT, and alignment scores in the synthetic task, suggesting that the proposed approach was well adapted to the synthetic data distribution. However, the lower performance on the standard caption prediction task indicates that the knowledge learned from synthetic captions did not fully transfer to real medical captions. This suggests the existence of a distribution gap between synthetic and real data, especially regarding factual correctness and clinical concept alignment.

Therefore, in conclusion, synthetic caption data can be useful for controlled model development, preliminary validation, and checkpoint selection, especially when access to real annotated medical images is limited. However, synthetic validation should not be considered a complete replacement for validation on real medical data. Future work should evaluate whether synthetic-validation rankings correlate with real-data rankings, combine synthetic and real captions during training, and incorporate factuality-aware objectives to improve clinical consistency and generalization.

For future work, several directions can be explored. First, a mixed training strategy combining synthetic and real caption data could be used to reduce the distribution gap observed in the official results. Instead of training exclusively on synthetic captions, the model could first be pretrained on synthetic data and then fine-tuned on real medical captions. This may allow the system to benefit from the scale and structure of synthetic data while adapting to the linguistic and clinical characteristics of real captions.

Second, future experiments should include stronger factuality-aware optimization. In the current

approach, the model was optimized using the standard causal language modeling objective and selected using a relevance-based validation score. However, the official results show that factuality remains one of the weakest aspects of the system, especially in the standard caption prediction task. Future work could incorporate medical concept supervision, entity-level losses, or reinforcement learning based on factuality metrics to encourage the generation of clinically consistent captions.

Finally, the architecture itself could be improved. Although the frozen CLIP encoder and Q-Former-style projector provide an efficient solution, medical-domain visual encoders or vision-language models pretrained on biomedical data may offer better representations for medical images. Replacing or adapting the visual backbone with a domain-specific encoder could improve the model's ability to identify anatomical structures, imaging modalities, and pathological findings.

Acknowledgments

This research is part of the research project MedicaLLM (CPP2024-011574) funded by MICIU/AEI /10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)-a way of making Europe.

Declaration on Generative AI

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

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[2] G. T. A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ram'e, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J.-B. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. I. 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. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. Gyorgy, 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. Z. 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-Pluci'nska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. M. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. 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, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. mei Xu, P. Stańczyk, P. D. Tafti, R. Shivanna, R. Wu, R. Pan, R. A. 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, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. S. 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, D. Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, ArXiv abs/2503.19786 (2025). URL: https://api.semanticscholar.org/CorpusID:277313563.

[3] G. Reale-Nosei, E. Amador-Domínguez, E. Serrano, From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation, Medical Image Analysis 97 (2024) 103264. URL: https://www.sciencedirect.com/science/article/pii/S1361841524001890. doi:https://doi.org/10.1016/j.media.2024.103264.

[4] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2497–2506.

[5] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), 2018, pp. 2577–2586.

[6] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 1439–1449.

[7] 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.

[8] 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.

[9] 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.

[10] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): A multimodal image dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, Cham, 2018, pp. 180–189.

[11] I. S. Ahmad, R. B. Suleiman, T. Yu, R. Lin, C. Zhao, J. Liao, B. Wu, Y. Xie, X. Liang, H. Wang, et al., Foundation models for x-ray interpretation: a narrative review of current techniques and future perspectives in diagnostic imaging, Quantitative Imaging in Medicine and Surgery 16 (2026) 321.

[12] M. Haggag, A. Amira, F. Kurugollu, H. Zaidi, B. Soudan, Automated diagnostic reporting from medical images using deep learning architectures: Methods, challenges, and future directions, Expert Systems with Applications (2026) 131946.

[13] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., Iclr 1 (2022) 3.

[15] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).