Textual Preprocessing and Multimodal Conditioning for Medical Image Captioning

Sebastián Rascón-Cervantes[1,*,†], Graciela Ramirez-Alonso[1,*,†], David R. Lopez-Flores[2,†] and A. Pastor López-Monroy[3,†]

[1]Comp. Vision & Data Science Lab, Universidad Autónoma de Chihuahua, Facultad de Ingeniería, 31125, Chihuahua, México
[2]Tecnológico Nacional de México / IT Chihuahua, Ave. Tecnológico 2909, Chihuahua, 31200, Chihuahua, México
[3]Department of Computer Science, Mathematics Research Center (CIMAT), Gto, México

Abstract

Medical image captioning aims to automatically generate clinically relevant textual descriptions from medical images, representing a challenging vision–language task due to the heterogeneity of imaging modalities, anatomical regions, and reporting styles. This paper presents the UACH-VisionLab team's submission to the ImageCLEFmedical Caption 2026 challenge. We propose a framework that combines textual preprocessing and multimodal conditioning to improve image–text alignment between medical images and their corresponding reports. The framework incorporates a two-stage LLM-based textual preprocessing pipeline consisting of cleaning and paraphrasing stages applied to the training captions, together with a multimodal captioning model that employs a frozen PubMedCLIP visual encoder and conditions a fine-tuned GPT-2 language model through continuous visual prefix embeddings. In the official ImageCLEFmedical Caption 2026 evaluation, the proposed system achieved an overall score of 0.2673, with a relevance score of 0.4275 and a factuality score of 0.1072. Qualitative examples illustrate the ability of the proposed framework to generate semantically coherent radiological descriptions across diverse imaging scenarios, while also highlighting persistent challenges related to factual consistency and anatomically precise grounding in heterogeneous medical image captioning tasks.

Keywords

Medical Image Captioning, Vision-Language Models, Image-Text Alignment, Text Preprocessing

1. Introduction

Medical imaging plays a central role in clinical diagnosis and treatment, leading to a continuous increase in the number of imaging studies performed each year. Consequently, the workload associated with radiological interpretation and report generation has grown substantially. Because medical reporting requires both visual analysis and the preparation of textual descriptions, the process is often time-consuming and represents a significant bottleneck in clinical practice [1, 2].

Recent advances in Vision-Language Models (VLMs) and Large Language Models (LLMs) have opened new possibilities for addressing these challenges by enabling multimodal systems capable of generating descriptive medical text directly from imaging data [3]. Such capabilities create opportunities to support clinical workflows by assisting in medical report generation and reducing the burden associated with radiological interpretation [4].

The ImageCLEFmedical Caption task [5], organized within the CLEF Initiative (Conference and Labs of the Evaluation Forum) [6], provides a heterogeneous benchmark composed of multiple imaging modalities and anatomical regions, creating a challenging evaluation scenario for medical image captioning systems. The diversity of modalities, anatomical structures, and reporting styles requires models to generalize across highly diverse visual patterns and clinical contexts. As part of the annual

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
fsebastian.rascon@gmail.com (S. Rascón-Cervantes); galonso@uach.mx (G. Ramirez-Alonso); david.lf@chihuahua.tecnm.mx (D. R. Lopez-Flores); pastor.lopez@cimat.mx (A. P. López-Monroy)
0009-0008-7641-314X (S. Rascón-Cervantes); 0000-0002-9781-3010 (G. Ramirez-Alonso); 0000-0003-4016-0845 (D. R. Lopez-Flores); 0000-0003-1018-4221 (A. P. López-Monroy)

challenge, participating teams submit their predictions through the AI4MediaBench platform, where systems are evaluated using six metrics covering both relevance and factuality aspects.

Recent approaches in medical image captioning have increasingly adopted Transformer-based vision–language architectures, commonly combining a visual encoder for image representation with a textual decoder for autoregressive caption generation [7]. More recently, several methods have incorporated intermediate mapping or alignment components designed to bridge the visual and textual embedding spaces, thereby improving image–text alignment and cross-modal conditioning during caption generation [8, 9].

Despite recent architectural advances, the quality of the training captions remains a critical factor in medical image captioning systems. During dataset inspection, it was observed that medical captions may contain non-visual information, acquisition details, patient-related metadata, or inconsistent reporting styles, which can negatively affect the correspondence between visual and textual information during multimodal training. To address these issues, this work builds upon the two-stage textual preprocessing pipeline previously presented in [10], where its methodology and experimental validation are described in detail. The resulting preprocessed captions are subsequently used to train the proposed multimodal captioning model.

This work presents the UACH-VisionLab team's approach for the ImageCLEFmedical Caption 2026 challenge. The proposed framework combines the preprocessed training corpus with a frozen PubMedCLIP visual encoder, a Transformer-based mapping network, and a fine-tuned GPT-2 language model for multimodal medical image captioning. Together, these components integrate textual preprocessing and multimodal conditioning to improve image–text alignment.

2. Dataset

For this work, the dataset provided for the ImageCLEFmedical Caption 2026 task was employed [5]. This edition corresponds to the 10th iteration of the ImageCLEFmedical Caption challenge and uses an updated and extended version of the ROCOv2 dataset [11]. The dataset is composed of radiology images sourced from biomedical articles available in the PubMed Central Open Access subset and includes curated image–caption pairs for medical image captioning.

The dataset covers multiple imaging modalities and anatomical regions, introducing substantial variability in both visual appearance and caption structure. Each image is associated with a textual caption describing radiological findings and anatomical information. The official dataset contains 97,364 training samples, 19,240 validation samples, and 15,249 test samples, with the test set consisting of previously unseen images used for challenge evaluation.

The dataset includes multiple imaging modalities, including Computed Tomography (CT), X-ray, Magnetic Resonance Imaging (MRI), Ultrasound, Angiography, Positron Emission Tomography (PET), PET/CT, and Optical Coherence Tomography (OCT). Among these modalities, CT and X-ray represent the largest proportion of images within the dataset.

A wide range of anatomical regions is also represented, particularly within X-ray studies, including the chest, cranium, lower extremities, abdomen, upper extremities, pelvis, spine, and breast regions. This anatomical diversity further increases the variability of both visual patterns and textual descriptions across the dataset.

Table 1 and Figure 1 illustrate the variability in caption length across the dataset, particularly within the training set employed for multimodal training and textual preprocessing. Although most captions are relatively short, the distribution presents a pronounced long-tail behavior, with substantially longer descriptions appearing in a smaller subset of samples. This variability reflects differences in caption structure and reporting style throughout the dataset.

Additionally, Figure 2 highlights the predominance of modality-related and anatomy-related terminology within the training captions. Frequent bigram patterns such as "computed tomography", "x ray", "magnetic resonance", and "t2 weighted" reveal the strong presence of modality-specific radiological language and structured reporting conventions throughout the dataset.

Table 1
Caption length statistics for the ImageCLEFmedical 2026 dataset splits.

Split	Mean	Median	Std	Min	Max
Train	21.34	17	16.6	1	767
Valid	22.73	18	19.0	0*	511

Note: Three validation captions produced a word count of zero during preprocessing because they were written in non-English language variants that were not properly detected by the employed tokenization procedure.

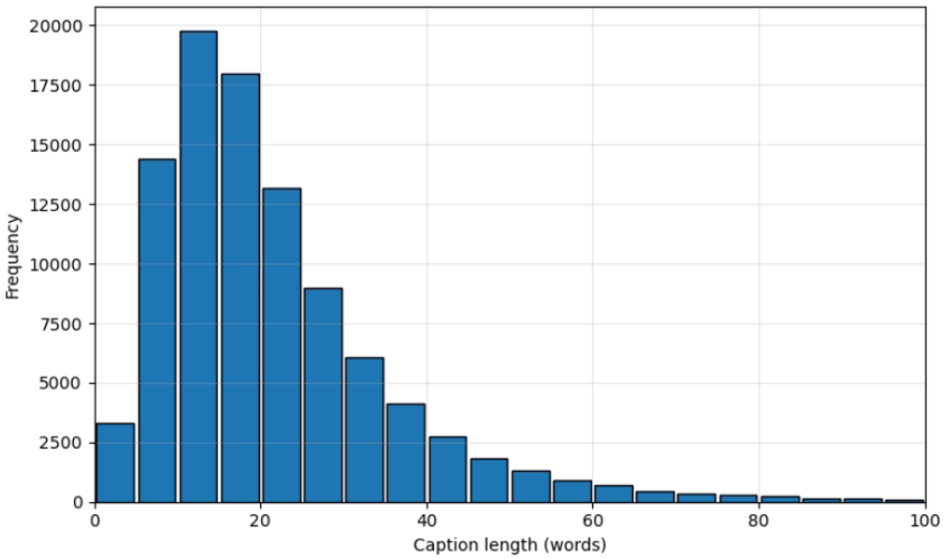


Figure 1: Caption length distribution for the training set.

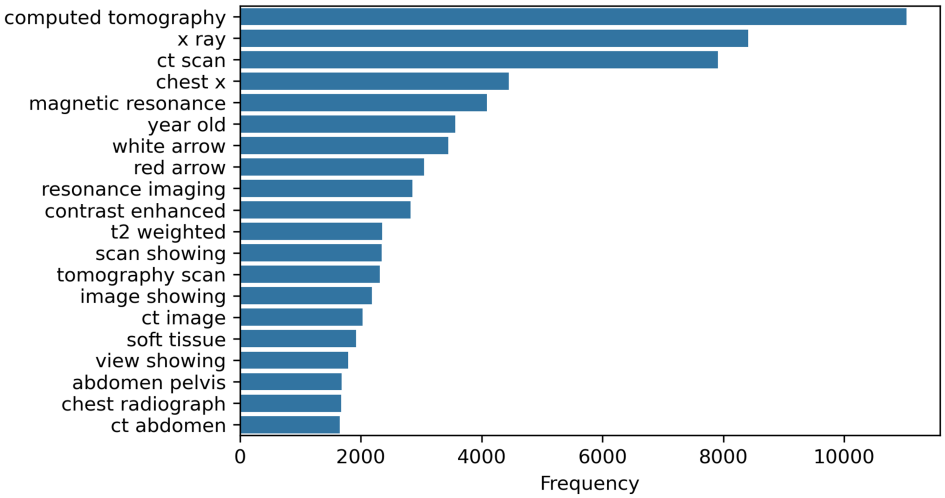


Figure 2: Top 20 most frequent bigrams after stop-word removal. Frequent bigram patterns reveal common modality- and anatomy-related terminology within the dataset captions.

3. Metrics

Following the official challenge evaluation protocol, model performance is assessed using six metrics covering both relevance and factuality aspects. Prior to evaluation, captions are converted to lowercase, punctuation is removed, and numerical values are replaced with a generic token.

Relevance is evaluated using Image-Caption Similarity [12], BERTScore Recall with inverse document

frequency (IDF) weighting [13], ROUGE-1 [14], and BLEURT [15]. Image-Caption Similarity measures semantic alignment between image and caption embeddings, while BERTScore evaluates contextual semantic similarity using BERT embeddings. ROUGE-1 computes unigram overlap, and BLEURT estimates text quality using a Transformer-based learned evaluation model.

Factuality is evaluated using UMLS Concept F1 and AlignScore [16]. UMLS Concept F1 is calculated following the methodology described in [17], where medical concepts are extracted using MedCAT and matched through Concept Unique Identifiers (CUIs), considering the semantic types employed by the MEDCON metric. AlignScore evaluates factual consistency between both texts through alignment-based scoring.

4. Related work

4.1. Medical Image Captioning and Vision–Language Models

Medical image captioning aims to automatically generate clinically relevant textual descriptions from medical imaging data by combining computer vision and natural language generation techniques. Early approaches commonly relied on encoder–decoder architectures, where convolutional neural networks (CNNs) were employed for visual feature extraction and recurrent neural networks (RNNs) were used for sequential text generation [2, 18].

More recent approaches have increasingly adopted Transformer-based architectures and Vision–Language Models (VLMs), enabling more effective multimodal representation learning and improved image–text interaction capabilities [3]. These models integrate visual encoders and autoregressive language decoders within unified multimodal frameworks capable of generating descriptive medical captions directly from imaging data [18].

4.2. Vision–Language Alignment through Contrastive Learning

Contrastive learning has emerged as an effective strategy for aligning visual and textual representations in Vision–Language Models. Contrastive Language–Image Pretraining (CLIP) learns a shared embedding space by maximizing the similarity between paired image–text samples while separating unrelated pairs, enabling strong multimodal representation learning across vision and language tasks [19].

However, general-domain contrastive models may present limitations when applied to specialized medical imaging scenarios due to domain differences between natural and medical images. To address this issue, several medical adaptations of CLIP have been proposed using medical image–text pairs for contrastive pretraining. Among them, PubMedCLIP [20] introduced a domain-specific contrastive framework trained on medical images and associated textual descriptions extracted from PubMed articles. Recently, BioMedCLIP [21] has advanced this area through large-scale biomedical image-text pretraining. Medical vision-language systems such as PCLMed [22] have used these specialized CLIP-based representations for medical caption generation and related multimodal tasks.

4.3. Prefix-Based Conditioning and Mapping Networks

Recent Vision–Language Models have incorporated intermediate mechanisms to improve interaction between visual and textual representations during autoregressive caption generation. Several multimodal captioning approaches have explored the use of mapping or projection modules that transform visual representations into continuous prompt embeddings capable of conditioning autoregressive language models through visual prefixing strategies [8, 23].

Other architectures such as BLIP-2 [9] introduced the Q-Former, a specialized intermediate module designed to facilitate information exchange between frozen visual encoders and large language models through learnable query representations. More recent medical captioning systems have adopted similar Q-Former-based strategies combined with partially trainable decoder components to improve multimodal interaction and image–text alignment during caption generation [24].

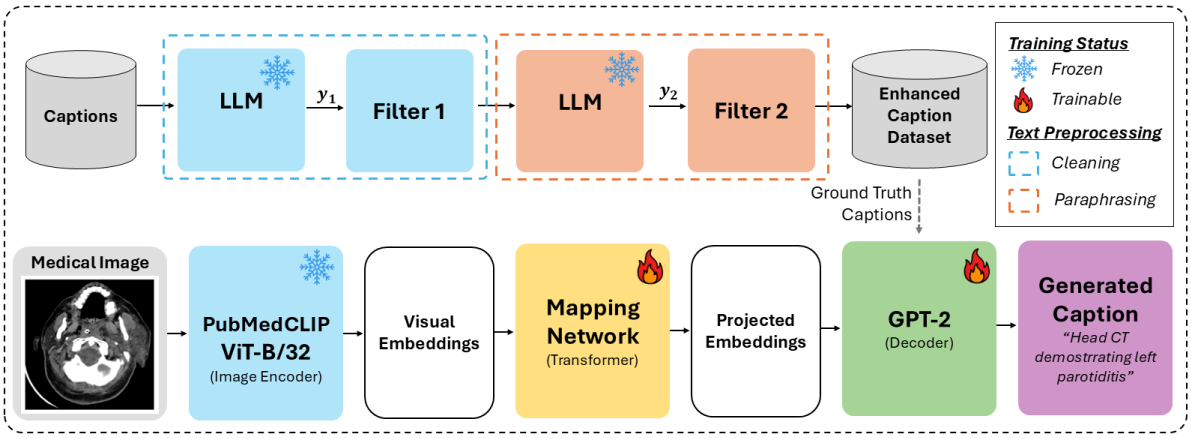


Figure 3: Overview of the proposed preprocessing pipeline and multimodal captioning architecture.

4.4. Caption Refinement and Textual Augmentation

Text preprocessing is a dataset-dependent process whose effectiveness relies on understanding the characteristics of the target data. Previous work has emphasized that preprocessing decisions should be guided by dataset exploration, since the most appropriate operations may vary according to the content and structure of each dataset [25].

LLMs have also been explored as tools for textual preprocessing and caption refinement. For example, prompt-based approaches have been used to remove textual noise and reduce caption length [26], while other works have used LLMs for data augmentation by generating conceptually related but semantically diverse samples [27, 28]. Chain-of-Thought prompting has also been adopted to guide the generation process during augmentation [29].

However, LLM-based preprocessing may introduce additional challenges, including incorrect or non-grounded outputs due to insufficient medical-domain specialization. These issues are related to hallucination phenomena such as factual incorrectness [30], motivating the use of controlled prompting and filtering mechanisms for reliable caption refinement. In clinical settings, lightweight LLMs have also been proposed for preprocessing under privacy and computational constraints, which are common considerations in medical applications [31].

Building upon the two-stage textual preprocessing pipeline previously presented in [10], this work integrates the previously proposed methodology into a multimodal medical image captioning framework based on a frozen PubMedCLIP visual encoder, a Transformer-based mapping network, and a fine-tuned GPT-2 language model. The proposed framework combines textual preprocessing with multimodal conditioning to address image–text alignment by using preprocessed training captions during model training and continuous visual prefix conditioning during caption generation. An overview of the complete preprocessing and multimodal framework is presented in Figure 3.

5. Methodology

This section describes the proposed framework for medical image caption generation. First, the textual preprocessing pipeline used to construct the enhanced caption dataset is presented. Then, the proposed multimodal captioning architecture based on PubMedCLIP, a transformer-based mapping network, and GPT-2 is described.

5.1. Text Preprocessing

The LLM-based preprocessing pipeline adopted in this work corresponds to the methodology previously presented in [10]. The resulting preprocessed corpus was used as the textual supervision for training

the proposed captioning model.

During dataset exploration, it was observed that several training captions contained non-visual or context-specific information, including patient metadata, temporal references, acquisition details, and procedural notes that could not be directly inferred from the corresponding image. Such non-visual information may introduce text–image misalignment in the supervision provided during multimodal training. To address this issue, the preprocessing pipeline consists of two sequential stages: (i) cleaning and (ii) paraphrasing.

During the cleaning stage, non-visual elements were removed while preserving clinically relevant findings that could be visually inferred from the image. A first filtering stage was incorporated after cleaning to discard captions containing insufficient visually grounded information, as well as captions shorter than five words that provided limited semantic context for reliable paraphrasing. Preliminary experiments showed that these short captions frequently caused the LLM to generate unnecessarily verbose or reasoning-like reformulations rather than concise caption variations.

Subsequently, the paraphrasing stage generated alternative textual formulations while preserving the original clinical meaning. After paraphrasing, a second filtering stage was applied to remove invalid, excessively long, or non-informative outputs generated by the LLM. Both preprocessing stages were implemented using `DeepSeek-R1-Distill-Llama-8B`. After the cleaning stage and the first filtering process, 88,544 valid captions were retained. These captions were subsequently paraphrased, yielding 71,356 additional caption variants after the second filtering stage. Consequently, the proposed cleaning-and-paraphrasing (C&P) pipeline generated a final training corpus containing 159,900 captions, which was used as ground-truth supervision during multimodal training.

Both preprocessing stages relied on structured prompting strategies to guide the LLM toward controlled caption refinement while minimizing verbose or non-grounded generations. Inspired by the structured prompting notation proposed in [32] and following prompt engineering best practices described in [33, 34, 35], each prompt was organized into four main components: an instruction, in-context examples, edge-case rules, and generation constraints.

The instruction component defined the role of the model as an expert radiologist and specified the objective of the corresponding preprocessing stage. In-context examples encouraged guided reasoning by providing representative input–output examples together with intermediate reasoning. Finally, edge-case rules and generation constraints were iteratively refined based on preliminary experiments to avoid invalid outputs and unnecessarily verbose generations.

The edge-case rules specified the expected behavior in uncommon situations. For example, during the cleaning stage, captions containing only non-visual information were mapped to “NA”, whereas captions without non-visual content were returned unchanged. During the paraphrasing stage, captions that were too short or difficult to reformulate without altering their semantic content were preserved. Additionally, generation constraints limited the reasoning process to a maximum of five sentences, since preliminary experiments showed that restricting only the output length was insufficient to prevent excessively verbose responses.

5.2. Proposed Multimodal Architecture

The proposed multimodal captioning architecture combines a frozen PubMedCLIP image encoder with a trainable Transformer-based mapping network and a fine-tuned GPT-2 language model.

Medical images are first encoded using PubMedCLIP ViT-B/32[1], a domain-specialized vision–language model pretrained on medical image–text pairs extracted from PubMed articles. Prior work [20] reported improved performance over general-domain CLIP representations in medical vision-language tasks, highlighting the benefits of medical-domain contrastive pretraining.

The PubMedCLIP encoder was kept frozen during training and used only as a visual feature extractor. Images were processed with the corresponding CLIP processor and encoded through the `get_image_features` function, producing a 512-dimensional visual embedding for each image. The

[1]The implementation used in this work corresponds to `flaviagiammarino/pubmed-clip-vit-base-patch32`.

resulting embedding was L2-normalized before being passed to the mapping network.

The mapping network [8] was implemented as an 8-layer Transformer-based projector that transforms PubMedCLIP visual representations into a sequence of 10 continuous prompt embeddings in the GPT-2 embedding space. During training, these continuous prompt embeddings are concatenated with the token embeddings of the target caption and provided as input to GPT-2.

These embeddings act as a visual prefix that conditions the language decoder during caption generation. During training, Gaussian noise with variance 0.016 was injected into the visual prefix as a regularization strategy. Finally, the pretrained `gpt2` checkpoint from Hugging Face, corresponding to the GPT-2 Small architecture (12 Transformer decoder layers, 12 attention heads, a hidden size of 768, a maximum context length of 1024 tokens, and approximately 124 million parameters), was fine-tuned to generate the target medical caption conditioned on the projected visual prefix. The model was optimized using AdamW with a learning rate of 2×10^{-5} for 4 epochs and a batch size of 1, while the PubMedCLIP visual encoder remained frozen throughout training.

5.3. Experimental Setup

All experiments were conducted on a workstation equipped with an Intel Core i9 processor, 32 GB of system RAM, and an NVIDIA TITAN RTX GPU with 24 GB of VRAM.

6. Results and Discussion

The proposed framework achieved an overall score of 0.2673 in the ImageCLEFmedical Caption 2026 challenge, obtaining a relevance score of 0.4275 and a factuality score of 0.1072. Official leaderboard results are summarized in Table 2.

Table 2
Leaderboard results for the ImageCLEFmedical Caption task. Boldface indicates the results obtained by the proposed framework.

Owner	Overall	Relevance	Factuality	BERT	ROUGE	Similarity	BLEURT	MedCAT	Align
AlstatLab	0.3755	0.5786	0.1724	0.6250	0.3234	1.0104	0.3555	0.2160	0.1287
UFPE-Essex-UNED	0.3603	0.5527	0.1679	0.6017	0.2725	0.9902	0.3463	0.1822	0.1537
dsgt_caption	0.3571	0.5365	0.1777	0.6042	0.2809	0.9381	0.3229	0.1943	0.1610
AUEB/DMST	0.3516	0.5288	0.1743	0.6130	0.2728	0.9113	0.3183	0.1812	0.1674
csmorgan	0.3458	0.5377	0.1539	0.6075	0.2725	0.9408	0.3299	0.1751	0.1327
gfreitas	0.3343	0.5201	0.1484	0.6013	0.2226	0.9273	0.3293	0.1563	0.1406
UIT_HighDefinition	0.3193	0.4999	0.1388	0.5977	0.2243	0.8496	0.3278	0.1397	0.1379
NUSTPB	0.3065	0.4705	0.1426	0.5465	0.2224	0.7787	0.3342	0.1443	0.1408
UMUTeam	0.2990	0.4798	0.1181	0.5670	0.2230	0.8147	0.3146	0.1048	0.1315
DS-MED Lab	0.2814	0.4526	0.1102	0.5690	0.2126	0.7286	0.3004	0.1427	0.0778
krishnatewari	0.2685	0.4346	0.1025	0.5398	0.2059	0.6788	0.3137	0.1219	0.0831
UACH-VisionLab	**0.2673**	**0.4275**	**0.1072**	**0.5742**	**0.1818**	**0.6589**	**0.2951**	**0.1155**	**0.0989**
nguyenthinh	0.2527	0.4373	0.0680	0.5720	0.2207	0.6514	0.3051	0.0872	0.0488

Overall, the leaderboard results reveal notable differences between relevance-oriented and factuality-oriented performance across participating systems. Metrics such as BERTScore and image–caption similarity achieve relatively high values for several approaches, suggesting that current Vision–Language Models are increasingly capable of generating semantically coherent descriptions with structurally consistent radiological language. However, factuality-oriented metrics, including MedCAT and AlignScore, remain substantially lower across most submissions, highlighting the persistent difficulty of preserving clinically precise findings and anatomically accurate details during caption generation.

These observations suggest that, although recent multimodal architectures can learn general radiological reporting patterns and semantic image–text associations, achieving robust clinical grounding and factual consistency remains a major challenge in heterogeneous medical captioning scenarios.

To further characterize the behavior of the proposed framework, an exploratory modality-level analysis was performed on the validation set. Validation samples with available imaging modality labels

were grouped according to their modality, and performance was assessed using BERTScore (F1) and ROUGE-L, two widely adopted text generation metrics. Samples without an available modality label were excluded from this analysis. The corresponding results are summarized in Table 3.

Table 3
Exploratory modality-level evaluation on the validation set using BERTScore (F1) and ROUGE-L. Only validation samples with available imaging modality labels are included.

Modality	N	BERTScore (F1)	ROUGE-L
CT	8234	0.6295	0.1674
X-ray	1511	0.7088	0.2635
MRI	1400	0.6548	0.1685
Ultrasound	686	0.6533	0.0845
PET	70	0.6305	0.1913

The modality-level analysis reveals noticeable performance differences across imaging modalities. The highest scores were obtained for X-ray images, achieving a BERTScore (F1) of 0.7088 and a ROUGE-L score of 0.2635. MRI, CT, and ultrasound exhibited relatively similar BERTScore values, whereas ultrasound obtained the lowest ROUGE-L score, indicating lower lexical overlap with the reference captions despite maintaining similar semantic similarity. PET results should be interpreted with caution due to the limited number of available samples. Overall, these findings suggest that caption generation performance varies across imaging modalities, motivating further investigation into modality-specific behavior in future work.

Qualitative examples from the validation set are presented in Table 4. The selected samples illustrate different levels of caption generation performance, ranging from semantically coherent radiological descriptions to partially preserved pathological findings and anatomical localization errors.

Examples 1 and 2 demonstrate cases where the proposed framework preserves clinically relevant findings and modality-specific radiological terminology. In Example 1, the model correctly identifies an abdominal X-ray study and generates a semantically coherent description consistent with bowel dilatation. Similarly, Example 2 shows clinically consistent terminology related to ground-glass opacities in chest CT imaging, although with reduced descriptive specificity compared to the reference caption.

In contrast, Examples 3 and 4 highlight limitations in preserving fine-grained pathological findings and precise anatomical grounding. In Example 3, the generated caption maintains the imaging modality and anatomical context while omitting the clinically important finding of severe external root resorption. Example 4 further illustrates partial semantic preservation combined with incorrect anatomical localization, where the model predicts a pancreatic lesion instead of a renal mass in abdominal CT imaging.

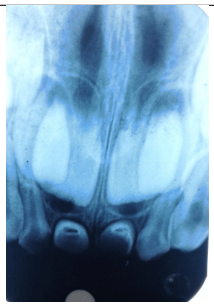
Overall, the qualitative examples illustrate that the proposed framework is capable of generating semantically coherent radiological descriptions while preserving modality-specific terminology and general anatomical context. Nevertheless, challenges related to fine-grained pathological findings, anatomically precise grounding, and factual consistency remain evident. These observations suggest that a more detailed quantitative analysis across imaging modalities and anatomical regions could provide further insights into the strengths and limitations of the proposed framework.

7. Conclusion

This work presented the UACH-VisionLab team's approach for the ImageCLEFmedical Caption 2026 challenge. The proposed framework integrates the previously presented two-stage LLM-based textual preprocessing pipeline with a frozen PubMedCLIP visual encoder, a Transformer-based mapping network, and a fine-tuned GPT-2 language model for multimodal medical image captioning.

Experimental results showed that the proposed approach was capable of generating semantically coherent radiological descriptions in several representative cases. The qualitative analysis further illustrated the model's ability to preserve clinically meaningful information in multiple imaging scenarios,

Table 4
Qualitative examples from the validation set. Blue text indicates modality or anatomical context, green text indicates clinically relevant information preserved in the generated caption, and red text indicates missing or incorrectly localized findings.

 CC BY [Jotham et al.] **Example 1:** **Target:** plane abdominal X-ray with a nonspecific dilatation of bowel loops **Prediction:** Abdominal X-ray demonstrates dilated small bowel loops.	 CC BY [Muacevic et al.] **Example 2:** **Target:** CT Chest with IV Contrast demonstrating ground-glass consolidative opacities throughout the bilateral lungs **Prediction:** HRCT thorax showing ground glass opacities in bilateral lung parenchyma
 CC BY [Muacevic et al.] **Example 3:** **Target:** Intraoral periapical radiograph showing the absence of roots due to severe external root resorption **Prediction:** Periapical radiograph of the maxillary left central incisor.	 CC BY [Shim et al.] **Example 4:** **Target:** Abdominal contrast-enhanced CT demonstrating a heterogeneous enhancing mass replacing the right kidney **Prediction:** Contrast-enhanced CT demonstrating a mass in the head of the pancreas

whereas the exploratory modality-level analysis revealed differences in caption generation performance across imaging modalities. Nevertheless, challenges related to factual consistency, fine-grained pathological findings, and anatomically precise grounding remain evident.

The obtained results highlight the persistent difficulty of medical image caption generation across diverse imaging modalities, anatomical regions, and reporting styles. Despite recent advances in Vision–Language Models, preserving clinically precise and factually grounded descriptions continues to represent an open research challenge.

Although the exploratory modality-level analysis provided initial insights into the behavior of the proposed framework across different imaging modalities, future work will investigate these modality-specific differences in greater depth using additional modality-specific and anatomical-region analyses. Additional research directions include improving factual consistency and anatomical grounding through retrieval-guided multimodal conditioning, alternative alignment mechanisms such as Q-Former architectures, and instruction-guided multimodal frameworks inspired by approaches such as InstructBLIP.

Acknowledgments

The authors acknowledge financial support from FECTI-IIC (Fondo Estatal de Ciencia, Tecnología e Innovación para el Estado de Chihuahua) through the project FECTI/2025/C02/IBA/019.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword, and content enhancement. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Online Resources

The source code of the proposed system is publicly available at: https://github.com/FrancescoSebastian/sebastian-imageclefmedical-caption-2026

References

[1] X. Hou, Z. Liu, X. Li, X. Li, S. Sang, Y. Zhang, MKCL: Medical knowledge with contrastive learning model for radiology report generation, Journal of Biomedical Informatics 146 (2023) 104496. doi:10.1016/j.jbi.2023.104496.

[2] X. Liu, J. Xin, Q. Shen, Z. Huang, Z. Wang, Automatic medical report generation based on deep learning: A state of the art survey, Computerized Medical Imaging and Graphics 120 (2025) 102486. doi:10.1016/j.compmedimag.2024.102486.

[3] M. Varol Arısoy, A. Arısoy, İ. Uysal, A vision attention driven language framework for medical report generation, Scientific Reports 15 (2025) 10704. doi:10.1038/s41598-025-95666-8.

[4] P. Nie, X. Liu, MedNet: A dual-copy mechanism for medical report generation from images, in: International Conference on Artificial Neural Networks, Springer, 2023, pp. 469–481.

[5] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryılmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026. To appear.

[6] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[7] 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 (2024) 103264.

[8] R. Mokady, A. Hertz, A. H. Bermano, ClipCap: CLIP prefix for image captioning, arXiv preprint arXiv:2111.09734 (2021).

[9] 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.

[10] S. Rascón-Cervantes, G. Ramirez-Alonso, A. P. López-Monroy, R. Lopez-Santillan, N. A. Rendón Mejía, A two-stage textual preprocessing pipeline for medical image captioning, in: Mexican Conference on Pattern Recognition, Springer, 2026, pp. 273–282.

[11] 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.

[12] N. C. Codella, Y. Jin, S. Jain, Y. Gu, H. H. Lee, A. B. Abacha, A. Santamaria-Pang, W. Guyman, N. Sangani, S. Zhang, et al., MedImageInsight: An open-source embedding model for general domain medical imaging, arXiv preprint arXiv:2410.06542 (2024).

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT (2020).

[14] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[15] 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.

[16] 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.

[17] 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.

[18] X. Liu, J. Xin, Q. Shen, Z. Huang, Z. Wang, Automatic medical report generation based on deep learning: A state of the art survey, Computerized Medical Imaging and Graphics 120 (2025) 102486. doi:10.1016/j.compmedimag.2024.102486.

[19] 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, 2021, pp. 8748–8763.

[20] S. Eslami, G. De Melo, C. Meinel, Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, arXiv preprint arXiv:2112.13906 (2021).

[21] 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.

[22] B. Yang, Y. Yu, Y. Zou, T. Zhang, PCLmed: Champion solution for ImageCLEFmedical 2024 caption prediction challenge via medical vision-language foundation models, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France, 2024.

[23] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, F. Zheng, Transferable decoding with visual entities for zero-shot image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3136–3146.

[24] Y. Lee, H. J. Kim, H. Shin, C. Lim, AI Stat Lab: A modular framework for clinically accurate medical image captioning using vision-language models, in: CEUR Workshop Proceedings, volume 4038, CEUR-WS, 2025, pp. 2511–2523.

[25] C. P. Chai, Comparison of text preprocessing methods, Natural language engineering 29 (2023) 509–553.

[26] E. Meguellati, N. Pratama, S. Sadiq, G. Demartini, Are large language models good data preprocessors?, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 2129–2132.

[27] Y. Chai, H. Xie, J. S. Qin, Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities, Artificial Intelligence Review 59 (2026) 35.

[28] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, F. Zeng, W. Liu, et al., AugGPT: Leveraging ChatGPT for text data augmentation, IEEE Transactions on Big Data (2025).

[29] H. Chen, Z. Dou, K. Mao, J. Liu, Z. Zhao, Generalizing conversational dense retrieval via llm-

cognition data augmentation, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2700–2718.

[30] S. Banerjee, A. Agarwal, S. Singla, LLMS will always hallucinate, and we need to live with this, in: Intelligent Systems Conference, Springer, 2025, pp. 624–648.

[31] Y. Qu, Y. Dai, S. Yu, P. Tanikella, M. Pillai, W. Chen, J. Xie, Y. Ren, D. Wang, Y. Wang, et al., PrecLLM: A privacy-preserving framework for efficient clinical annotation extraction from unstructured ehrs using small-scale LLMs, Research Square (2026) rs–3.

[32] X. Liang, D. Wang, H. Zhong, Q. Wang, R. Li, R. Jia, B. Wan, Candidate-heuristic in-context learning: A new framework for enhancing medical visual question answering with llms, Information Processing & Management 61 (2024) 103805.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[34] P. Shojaee, I. Mirzadeh, M. Horton, S. Bengio, M. Farajtabar, et al., The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, Advances in Neural Information Processing Systems 38 (2026) 108018–108059.

[35] Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, N. Zou, et al., Stop overthinking: A survey on efficient reasoning for large language models, Trans. Mach. Learn. Res. (2025).