AI Stat Lab: A Modular Framework for Clinically Accurate Medical Image Captioning Using Vision-Language Models

Notebook for the ImageCLEF Lab at CLEF 2025

Yunseo Lee[1,†], Hyun Jun Kim[2,†], Heeseung Shin[1,†] and Changwon Lim[3,*]

[1]Department of Statistics and Data Science, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea
[2]Department of Smart Cities, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea
[3]Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

Abstract

We propose a modular framework for medical image captioning that integrates domain-adapted visual encoders, token-efficient representation via query-based compression, and post-hoc refinement. The architecture employs an ensemble of general-purpose and domain-specific vision encoders (SigLIP2 and BioMedCLIP), a Q-Former for dense concept-aware tokenization, and a LoRA-tuned Bio-Medical LLaMA-3 decoder. Auxiliary objectives guide the model to jointly predict UMLS concepts and semantic types, improving semantic grounding. At inference, captions from six independently trained variants are reranked using three complementary strategies—BioMedCLIP similarity, BLEURT scoring, and BioBERT-based centroid alignment. Evaluations on the ImageCLEF2025 Caption Prediction Task demonstrate consistent gains in semantic relevance and clinical factuality over single-encoder and non-multitask baselines. Our approach (team: AI Stat Lab, ID #1900) achieved third place with an overall score of 0.3229, corresponding to relevance and factuality scores of 0.5089 and 0.1369, respectively.

Keywords

Medical image captioning, Vision-language model, Dual Encoder, UMLS concepts, Caption reranking, GPT summarization,

1. Introduction

Medical image captioning, automatically generating radiologist-style descriptions from imaging studies, has the potential to accelerate report drafting, improve content-based image retrieval, and increase the interpretability of diagnostic AI models. Compared with natural-image captioning, the task is complicated by grayscale modalities, subtle anatomical cues, and a highly specialized vocabulary, all of which demand fine-grained visual reasoning and domain knowledge [1].

While prior efforts have made notable progress by employing encoder-decoder frameworks trained on paired image-text datasets, the performance of these systems is often hindered by limitations in data quality, domain adaptability, and output reliability [2, 3, 4]. For instance, low-resolution images [5] and annotation-induced artifacts are prevalent in public medical datasets [6], degrading model perception. Moreover, generic vision encoders may lack the capacity to extract subtle domain-specific features [7], and caption decoders often produce inconsistent or incomplete descriptions due to limited grounding in clinical semantics [8]. To address these limitations, we construct a modular medical captioning framework by assembling and adapting proven techniques across the visual and language modeling pipeline. In particular, the pre-processing stage includes resolution enhancement and visual consistency adjustments [9, 10]. To address these limitations, we construct a modular medical captioning framework by assembling and adapting proven techniques across the visual and language modeling pipeline. Specifically, we integrate:

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
34_04@naver.com (Y. Lee); hyunjun0615@cau.ac.kr (H. J. Kim); hs970416@cau.ac.kr (H. Shin); clim@cau.ac.kr (C. Lim)
0009-0005-0390-3541 (Y. Lee); 0009-0000-7512-7987 (H. J. Kim); 0009-0004-0430-9547 (H. Shin); 0000-0002-3182-8420 (C. Lim)

1. A dual-encoder configuration using SigLIP2 [11] and BioMedCLIP [12] for both general and domain-specific feature extraction [13],
2. A Query Transformer (Q-Former) [14] to reduce redundancy and enable concept-aware representations,
3. A biomedical LLaMA-3 decoder [15] fine-tuned via Low-Rank Adaptation (LoRA) for efficient adaptation, and
4. A post-hoc refinement stage that consolidates outputs from six independently trained captioning models.

This module employs GPT-4-based summarization [16] and multiple reranking strategies—including BiomedCLIP similarity, BLEURT [17] scoring, and centroid-based selection [18] using BioBERT [19]—to generate a single, clinically coherent caption.

2. Related Works

Medical image captioning has evolved alongside advances in vision-language modeling, primarily following the encoder-decoder paradigm widely used in natural image captioning. Early works employed convolutional neural networks (CNNs) as visual encoders paired with recurrent neural networks (RNNs) or Transformer-based decoders to generate captions [1]. However, these approaches often lacked clinical specificity, as they relied on general-purpose image features and were trained on limited or noisy medical datasets.

More recently, the integration of large-scale vision-language models (VLMs), such as BioMedCLIP [12], has enabled more transferable and semantically rich representations across diverse medical imaging modalities [19, 20, 21, 22]. These models, pretrained on multimodal datasets, facilitate improved generalization to unseen clinical data with minimal supervision.

Furthermore, the advent of large language models (LLMs), including GPT [23] and LLaMA [24], has further advanced captioning performance by providing enhanced language fluency, contextual reasoning, and factual alignment. Some recent systems incorporate LLMs as decoders conditioned on image-derived embeddings or prompts, allowing for richer and more coherent textual outputs.

In parallel, post-hoc refinement strategies have emerged as a practical solution for improving caption consistency. Ensemble-based generation followed by reranking using clinical relevance metrics—such as BERTScore [25], BLEURT [17], and visual-semantic similarity—has shown promise in reducing redundancy and hallucination. GPT-based summarization has also been explored to consolidate conflicting candidate captions into a single coherent report.

3. Method

3.1. Pre-processing

Training images in ROCOv2 exhibit two systematic defects, low spatial resolution and bright border artifacts, that degrade visual embeddings and, by extension, caption quality. We therefore apply a two-stage pre-processing pipeline comprising super-resolution and structure-aware inpainting.

First, we observed that 3,485 training images exhibited spatial resolutions smaller than 300 × 300 pixels. Considering the non-negligible proportion of such images and the risk of losing fine-grained visual cues crucial for captioning, we applied 2× super-resolution to these samples. For this purpose, we utilized the Feedback Adaptive Weighted Dense Network (FAWDN) [9] a recurrent convolutional architecture equipped with a feedback mechanism and adaptive dense blocks. FAWDN progressively refines image quality over multiple time steps by combining current inputs with hidden states from previous iterations. The network is composed of shared input, hidden, and output units across all time steps, and integrates an Adaptive Weighted Dense Block (AWDB) that captures multi-scale features through a combination of 1×1 convolutional layers and dense connections. This network was selected

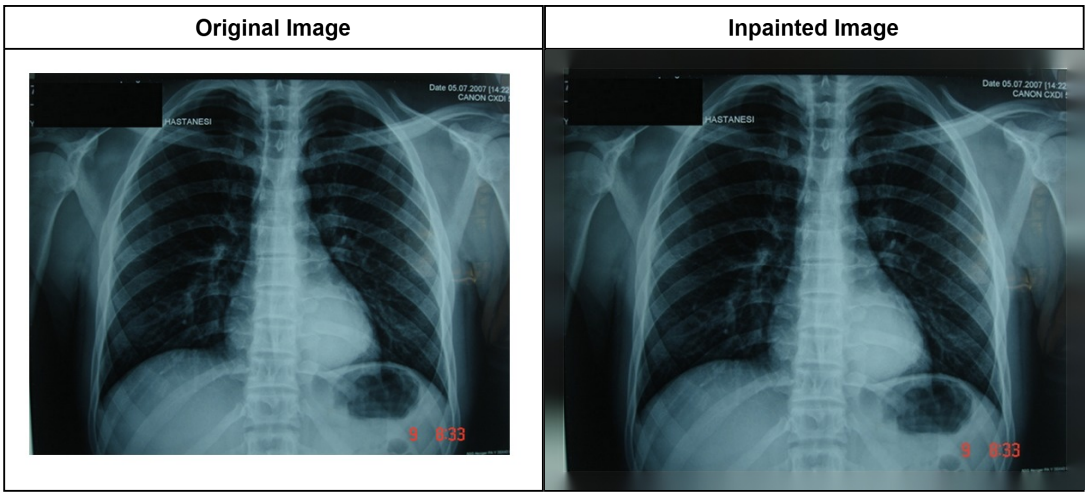

Figure 1: Example of structure-aware inpainting applied to bright borders. The method effectively removes overly bright edge regions while preserving structural consistency in the image.; CC BY [Seyhan et al.] [26]

not only for its proven performance on diverse image datasets but also due to the availability of pre-trained models specific to the medical domain, allowing us to avoid resource-intensive training of super-resolution models from scratch.

Second, to address the frequent presence of white or overly bright borders in the dataset images—often resulting from scanning artifacts or annotation overlays—we implemented a structure-aware inpainting strategy instead of simple cropping. Specifically, we identified border regions with brightness levels exceeding 245 within a fixed 8% margin around the image edges and applied the inpainting algorithm introduced by Telea [10] to fill these regions using nearby pixel information. Representative results are shown in Figure 1. Unlike hard cropping, which risks discarding medically relevant content near the periphery, this inpainting method preserves the overall anatomical integrity of each image while eliminating non-informative border artifacts. This procedure enhances the visual consistency of inputs and prevents the model from learning spurious cues unrelated to the actual medical content.

Together, these pre-processing steps improve the signal-to-noise ratio in the image encoder input and help stabilize caption generation by standardizing input quality across the dataset.

3.2. Model Architecture

The overall architecture of our proposed medical image captioning model is illustrated in Figure 2. The model consists of dual vision encoders, a Query Transformer (Q-Former), and a domain-adapted LLaMA decoder, which are described in detail in the following subsections.

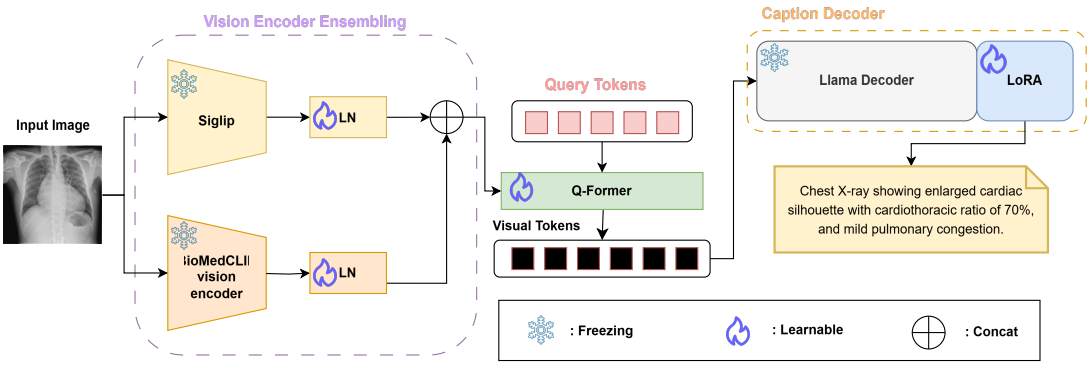

Figure 2: The figure illustrates the architecture of a medical image captioning model that generates a final caption by fusing outputs from two vision encoders, followed by a Q-Former and a LLaMA decoder.; CC BY-NC [Al Mulhim et al.] [27]

3.2.1. Dual Encoder

To derive robust and semantically rich visual representations from medical images, we adopt an ensemble of two vision encoders. Specifically, we utilize SigLIP2 [11], a general-purpose image encoder pretrained on large-scale natural image-text pairs, and BioMedCLIP [12], a medical-domain-specific encoder trained on 15 million image-caption pairs mined from PubMed Central.

To address the lack of medical image knowledge in the original SigLIP2 model, we perform domain-specific pre-adaptation by fine-tuning it on the dataset provided by the ImageCLEF2025 Caption Prediction Task [28, 29]. This enhances the encoder's ability to capture domain-relevant visual features while preserving generalization capacity. Following standard practice, we remove the classification heads from both encoders and extract intermediate features from their penultimate transformer layers. Let the feature outputs from BioMedCLIP and SigLIP2 be denoted as $\mathbf{f}_{\text{bioclip}} \in \mathbb{R}^{B\times 768}$ and $\mathbf{f}_{\text{siglip}} \in \mathbb{R}^{B\times 1536}$, respectively. These representations are concatenated to form a unified embedding $\mathbf{f} = [\mathbf{f}_{\text{bioclip}}; \mathbf{f}_{\text{siglip}}] \in \mathbb{R}^{B\times 2304}$, which preserves domain-specific detail from BioMedCLIP and high-level semantics from SigLIP2.

3.2.2. Query Transformer (Q-Former)

To reduce redundancy and computational burden, we apply a Q-Former [14] that projects the high-dimensional visual embedding $\mathbf{f} \in \mathbb{R}^{B\times 2304}$ into a fixed number of informative latent tokens. The visual feature $\mathbf{f}$ is broadcast across a learnable set of query tokens, resulting in a sequence input $\mathbf{X} \in \mathbb{R}^{B\times 32\times 2304}$. The Q-Former consists of six transformer layers with cross-attention modules that allow each query token to selectively attend to parts of the visual input. The output of the Q-Former is denoted as $\mathbf{Z} \in \mathbb{R}^{B\times 32\times 4096}$. This output is used for both caption generation and auxiliary concept classification.

To enhance medical grounding, we incorporate a multitask classification objective [30]. The output $\mathbf{Z} \in \mathbb{R}^{B\times 32\times 4096}$ is mean-pooled across the query dimension to produce a global representation $\bar{\mathbf{z}} \in \mathbb{R}^{B\times 4096}$. This representation is passed through two linear classifiers: one to predict concept presence among 2,478 Concept Unique Identifiers (CUIs), and another to predict 21 coarse concept types. The overall loss function is a weighted combination of the captioning loss and classification loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{caption}} + \lambda \cdot \mathcal{L}_{\text{cls}}$$

where $\mathcal{L}_{\text{caption}}$ is the cross-entropy loss over caption tokens, and the auxiliary term $\mathcal{L}_{\text{cls}}$ uses the multi-label margin loss. This multi-task setup improves alignment between the generated captions and clinical concepts visually present in the input image.

3.2.3. Caption Decoder

For the caption generation task, we adopt Bio-Medical LLaMA-3-8B [15] a domain-specialized variant of Meta-Llama-3-8B-Instruct [31] as the language decoder. The model has been fine-tuned on BioMedData, a high-quality biomedical dataset containing over 500,000 entries. The dataset comprises a blend of synthetic and manually curated samples, enabling robust generalization across a wide range of biomedical contexts. During training, the 32 Q-Former tokens are inserted as prefix embeddings that condition every decoding step on visual evidence. To enable efficient fine-tuning, we incorporate LoRA [32] modules into the decoder. This allows the model to adapt to medical image captioning tasks with minimal parameter updates while preserving the core language modeling capabilities of LLaMA.

3.2.4. Model Variants for Ensemble

To improve caption diversity and stabilize final output quality, we trained six independently parameterized captioning models under varying training configurations. All models share the same core architecture consisting of a Q-Former module and a LLaMA-based language decoder, but differ in their visual encoder types and auxiliary training settings. Specifically, we constructed two models each for

three encoder configurations: (1) using BioMedCLIP alone, (2) using SigLIP2 alone, and (3) using a dual-encoder setup that concatenates both BioMedCLIP and SigLIP2. Within each encoder group, one model was trained with auxiliary concept classification (predicting UMLS concepts and types [33]) and one without it. These six models generate diverse caption candidates for each image, forming the foundation for our post-processing pipeline described in the next subsection.

3.3. Post-processing

To further refine the raw captions generated by our six independently trained captioning models, we applied post-processing strategies aimed at improving both clinical coherence and factual relevance. This section presents two major post-processing components: (1) summarization-based refinement using GPT APIs and (2) candidate caption reranking based on semantic and domain-specific metrics.

3.3.1. Summarization-based Refinement

We employed two GPT-4-based summarization [16] strategies to consolidate the six candidate captions—each produced by a different model—into a single, medically accurate sentence. Both approaches aimed to improve readability, reduce redundancy, and ensure consistency with structured medical knowledge. The exact prompts used for each summarization method are provided in Table 1 below.

Table 1
Two distinct prompting strategies for GPT-4-based medical image caption summarization are presented. The Chain-of-Thought approach decomposes each input caption into key semantic components (e.g., modality, anatomical location, and pathological findings) and aggregates them through token-level consensus. In contrast, the Prompt-guided approach directly synthesizes multiple captions into a clinically accurate and coherent single-sentence summary. Both methods enable GPT-4 to generate standardized radiology reports from diverse input descriptions.

Chain-of-Thought Summarization template	**Prompt-guided Summarization template**
You are a board-certified radiologist. **TASK** 1. Parse EACH caption and list by line: <MODALITY>, <ANATOMIC_SITE>, <PATHOLOGIES>, etc. 2. Build a CONSENSUS table of token frequency. 3. Resolve conflicts by majority vote or keep the longer/specific one. 4. Compose ONE radiology-style sentence (35–45 words): – retains exact terms from table – concatenates: modality → site → key findings → clinical context – uses “shows”, “demonstrates”, avoids headings – omits absent content. OUTPUT: FINAL_CAPTION: <your summary> CAPTIONS: • caption 1: {caption1} • caption 2: {caption2} • caption 3: {caption3} • caption 4: {caption4} • caption 5: {caption5} • caption 6: {caption6}	You are a radiologist summarizing multiple captions of a medical image into ONE detailed sentence. – Integrate the imaging modality, anatomical location, pathological findings, and specific clinical details. – Use medically correct, extractive phrasing that maximizes token overlap—avoid paraphrasing unless synonymous medical terminology improves clarity. – Use present continuous tense with a subject-predicate-object structure. – Keep the summary natural, clinically accurate, and around 40 words, allowing slight variation if shorter or longer improves clarity. – If captions contain inconsistencies, prioritize findings with the highest diagnostic or therapeutic relevance. HERE are the captions: • caption 1: {caption1} • caption 2: {caption2} • caption 3: {caption3} • caption 4: {caption4} • caption 5: {caption5} • caption 6: {caption6}

Prompt-guided Summarization From each captioning models, six caption candidates were aggregated and fed into a standardized GPT-4 prompt. The prompt requested a concise and clinically coherent summary under the assumption that these captions describe the same medical image. This helped filter out redundant or inconsistent information and unify expression styles across captions.

Chain-of-Thought Summarization In this variant, the prompt instructed the model to generate step-by-step reasoning [34] before concluding the final summary. The intent was to increase factual consistency by encouraging the model to align each summary point with the underlying clinical evidence extracted from input captions. Empirically, this strategy improved alignment with structured medical entities.

3.3.2. Caption Reranking

To select the most appropriate caption among the generated candidates, we implemented a reranking module based on three different metrics: BioMedCLIP-image-text alignment, BLEURT-self-consensus, BioBERT centroid proximity. The overall framework of these reranking strategies is illustrated in Figure 3.

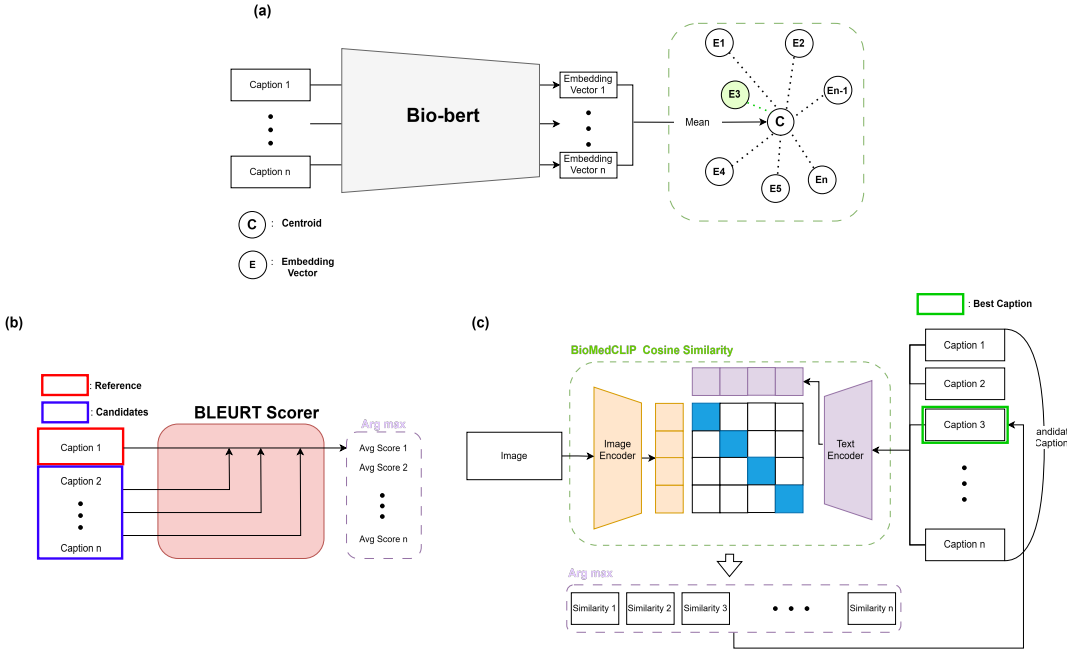

Figure 3: Overview of reranking methods: (a) BioBERT centroid similarity, (b) BLEURT self-consensus, (c) BioMedCLIP image-text alignment

BioMedCLIP-image-text alignment Each caption c_i is embedded into v_i with the BiomedCLIP [13] text encoder, and the corresponding image I is embedded into w using the image encoder. Cosine similarity

$$\text{sim}_i = \cos(\mathbf{v}_i, \mathbf{w}) = \frac{\mathbf{v}_i \cdot \mathbf{w}}{\|\mathbf{v}_i\| \cdot \|\mathbf{w}\|}$$

measures visual-textual coherence in a biomedical semantic space. The caption with the highest similarity score is selected:

$$\hat{c} = \arg\max_i \text{sim}_i$$

BLEURT-self-consensus BLEURT [17] estimates sentence quality via a regression head over BERT-style embeddings. For caption c_i among n candidates, we compute the leave-one-out average

$$\text{score}_i = \frac{1}{n-1} \sum_{j \neq i} \text{BLEURT}(c_i, c_j)$$

which rewards captions that are semantically central to the hypothesis set and thus robust to outliers. The final caption is selected by maximizing this self-consistency score:

$$\hat{c} = \arg\max_i \text{score}_i$$

BioBERT centroid proximity All captions are embedded via BioBERT [19] as vectors $v_1, \ldots, v_n$. The centroid

$$\mathbf{v}_c = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_i$$

represents the consensus sematic position. Each caption is then ranked based on Euclidean distance to the centroid [18].

$$\hat{c} = \arg\min_i \|\mathbf{v}_i - \mathbf{v}_c\|$$

4. Experiments

4.1. Experimental Setups

Dataset We conduct experiments on the extended version of the ROCOv2 dataset[28, 35], specifically curated for the ImageCLEFmedical 2025 Caption Prediction Task [28]. Unlike the original ROCOv2 [35], this updated release includes additional manual annotations as well as a newly introduced test set for the 2025 challenge. The dataset configuration differs from prior versions: the previous test set from ROCOv2 has been reassigned as the validation set, and the prior validation set has been merged into the training set. The newly collected 2025 test set contains unseen images to evaluate generalization performance under updated task conditions. The resulting splits comprise 80,091 images for training, 17,277 for validation, and 19,267 for testing. Each image is associated with a manually curated caption and UMLS concepts, making it suitable for both generation and concept detection tasks.

Evaluation Metrics Model performance is evaluated according to the official challenge protocol using six metrics that assess both relevance and factuality. Relevance is assessed using BERTScore (Recall with IDF), ROUGE-1 (F1), BLEURT and Image-text Similarity, with BERTScore is computed with the microsoft/deberta-xlarge-mnli model using IDF scores derived from the test set and BLEURT using the recommended BLEURT-20 checkpoint. Image-text Similarity is evaluated by independently extracting embedding vectors for the image and its corresponding caption using the MedImageInsight [36] model, followed by computing their cosine similarity. All relevance metrics are calculated on lowercase, punctuation-free captions with numbers replaced by the token "number." For factuality, UMLS Concept F1 is computed using MedCAT and semantic type filtering via QuickUMLS, and AlignScore is used to measure information consistency between predicted and reference captions based on RoBERTa-base alignment. All scores are averaged over the entire test corpus.

Model Settings Our system utilizes either BioMedCLIP or SigLIP2 as standalone vision encoders, or their ensemble via channel-wise feature concatenation. The language decoder is Bio-Medical LLaMA-3-8B, a domain-specific large language model. Visual features are processed by a 6-layer Q-Former with 32 learnable query tokens. The Q-Former maps from a 2304-dim input (concatenated encoder outputs) to 4096-dim embeddings compatible with the LLM. The model optionally includes auxiliary classification heads that predict 2,478 UMLS concept labels and 21 coarse types. The total loss is computed as a weighted sum of the captioning loss and the concept classification loss, where the weighting factor λ is empirically set to 0.1.

The model was trained using the AdamW optimizer, with the learning rate linearly increased to 1e-4 during the first epoch and annealed to 1e-6 over a total of 10 epochs. Training was conducted on a NVIDIA H100 GPU with a batch size of 16 and a gradient accumulation step of 2. During inference, we

employed beam search decoding with a beam width of 3, a repetition penalty of 2.5, a length penalty of 2.0, and a minimum and maximum output length of 8 and 64 tokens, respectively.

Image and Text Pre-processing To mitigate quality degradation caused by low-resolution inputs (<300×300) and overly bright borders, we implemented a two-stage pre-processing pipeline comprising FAWDN-based 2× super-resolution and structure-aware inpainting described in 3.1. In addition, we experimented with applying a Gaussian filter during image pre-processing and GPT-based back-translation [37] (English → Korean → English) for text augmentation; however, neither approach yielded notable performance improvements.

4.2. Base Model Results

Table 2
Summarizes the performance of six backbone configurations that differ in (i) vision-encoder composition and (ii) presence of auxiliary concept heads. All models share the Q-Former + BioMedLLAMA-3-8B decoding stack and are trained under the identical hyper-parameter schedule. Bold numbers indicate the column best on the test split; † denotes the best within the valid split.

Vision Encoder	Aux. CUI/type	Split	Image-text SIM	BERTScore (Recall)	ROUGE-1	BLEURT	ALIGN	UMLS F1
BioMedCLIP (#1405)	X	Valid	-	0.5845	0.2261	0.3100	0.1086	0.1405
		Test	0.8223	0.5756	0.2279	0.3094	0.1117	0.1382
SigLIP2 (#1407)	X	Valid	-	0.5796	0.2194	0.3047	0.0962	0.1397
		Test	0.8273	0.5710	0.2221	0.3049	0.1008	0.1320
Dual Encoder (#1673)	X	Valid	-	0.5826	0.2305	0.3133	0.1162†	0.1514
		Test	0.8365	0.5741	0.2328	0.3130	**0.1220**	0.1439
BioMedCLIP (#1693)	O	Valid	-	0.5860	0.2252	0.3100	0.1077	0.1411
		Test	0.8203	**0.5777**	0.2270	0.3094	0.1101	0.1403
SigLIP2 (#1694)	O	Valid	-	0.5837	0.2289	0.3090	0.0321	0.1438
		Test	-	-	-	-	-	-
Dual Encoder (#1695)	O	Valid	-	0.5863†	0.2347†	0.3150†	0.1148	0.1528†
		Test	**0.8491**	**0.5775**	**0.2390**	**0.3154**	0.1167	**0.1487**

We conducted ablation experiments to evaluate the impact of the dual encoder architecture and auxiliary classification tasks on medical image captioning performance. The detailed performance comparison across model variants is presented in Table 2. Compared to single-encoder baselines using either BioMedCLIP (#1405) or SigLIP2 (#1407), the dual encoder model (#1673), which concatenates both encoders along the channel dimension, consistently outperformed in terms of both relevance and factual accuracy. On the test set, this model improved ROUGE-1 and BLEURT scores by up to +0.0107 and +0.0081, respectively, while UMLS F1 increased by as much as +0.0119, demonstrating the effectiveness of combining domain-specific and general-purpose visual representations.

Building upon this, we introduced auxiliary classification heads for predicting UMLS concepts and semantic types. Compared to the base dual encoder (#1673), the model with concept prediction (#1695) achieved further gains across all major metrics, including an additional +0.0062 in ROUGE-1 and +0.0048 in UMLS F1. These improvements underscore the value of explicitly modeling medical concepts, which enhances the factual grounding of generated captions without sacrificing fluency. Among all configurations, the dual encoder with auxiliary classification (#1695) achieved the strongest overall performance, ranking first in four out of six evaluation metrics. These findings validate our architectural

choices: integrating heterogeneous visual features through a dual encoder and reinforcing clinical relevance through concept-aware auxiliary tasks. Together, these components contribute to generating more accurate, informative, and clinically coherent medical image captions.

4.3. Post-Processing Results

Table 3
Contrasts three reranking strategies applied to the pool of six captions produced per image. All methods outperform the strongest base model (#1695; Table 2), underscoring the value of hypothesis selection after decoding. For each metric, the best score on the test split is highlighted in bold. (* includes GPT-4 Chain-of-Thought and prompt-guided summaries among the candidates.)

Sub-ID	Reranker	Split	Image-text SIM	BERTScore (Recall)	ROUGE-1	BLEURT	ALIGN	UMLS F1
#1900	BioMedCLIP	Valid	–	0.5873	0.2338	0.3130	0.1095	0.1499
		Test	0.8919	0.5823	**0.2440**	0.3173	0.1231	0.1524
#1965	BLEURT*	Valid	–	0.5880	0.2368	0.3178	0.1221	0.1539
		Test	**0.9008**	0.5813	0.2397	**0.3186**	0.1162	0.1486
#1944	bioBERT	Valid	–	0.5922	0.2409	0.3179	0.1202	0.1552
		Test	0.8510	**0.5854**	0.2437	0.3178	**0.1233**	**0.1536**

We evaluate three caption reranking strategies, namely BioMedCLIP-base, BLEURT-base, and bioBERT-base, by selecting the best caption among candidates generated from six base models. All three reranking methods improve upon the base model outputs, confirming that post-processing plays a critical role in enhancing caption quality. Among the methods, BLEURT-based reranking (#1965) achieved the highest cosine similarity (0.9008) and BLEURT score (0.3186), while maintaining strong results across ROUGE (0.2397) and UMLS F1 (0.1486). This suggests that selecting internally consistent captions—those that align with the majority of candidate hypotheses—enhances both fluency and factual alignment.

BioMedCLIP-based reranking (#1900) prioritized visual-semantic grounding, yielding the highest ROUGE (0.2440) and competitive scores in ALIGN (0.1231) and UMLS F1 (0.1524). In contrast, BioBERT-based reranking (#1944) produced the highest BERTScore (0.5854), along with balanced performance across all metrics and the strongest UMLS F1 (0.1536). These results demonstrate that reranking not only improves overall caption quality but also enables fine-grained control depending on whether fluency, alignment, or clinical factuality is prioritized.

In addition to reranking methods, GPT-4-guided summarization was assessed as an alternative post-processing strategy. Despite the intuitive appeal of aggregating multiple candidate captions into a single concise output, these summarization strategies underperformed relative to reranking in our quantitative assessments. Specifically, both the prompt-guided and chain-of-thought prompting approaches frequently exhibited reduced precision and occasionally introduced clinically irrelevant or hallucinated content. These shortcomings were particularly evident in factual grounding metrics such as ALIGN and UMLS F1, indicating that generative summarization may abstract away or omit key clinical entities during compression. A detailed comparison of these limitations is provided in Table 4.

In summary, each reranking strategy exhibits distinct advantages: BLEURT-base excels in fluency and self-consistency, BioMedCLIP-base in vision-language alignment, and bioBERT-base in semantic grounding and metric balance. These results highlight that the choice of reranking method should depend on the specific priorities of the medical captioning application. Given that both relevance and factuality were key evaluation metrics in the ImageCLEF2025 challenge, the post-processing approach based on image-text alignment using BioMedCLIP achieved the highest performance. This submission, made under the team name AI Stat Lab with submission ID #1900, achieved an overall score of 0.3229 and ranked third on the official leaderboard.

5. Conclusion

We introduced a modular medical-image captioning pipeline that unifies three components: (i) dual vision encoders (BioMedCLIP + SigLIP2) to fuse domain-specific and general visual knowledge, (ii) a multitask loss that aligns captions with 2,478 UMLS concepts and 21 semantic types, and (iii) a metric-aware reranker that selects the most faithful hypothesis among six candidates. Specifically, our best submission (#1900) was constructed by applying BioMedCLIP-based reranking to the pool of six candidates generated from six base models (submissions #1405, #1407, #1673, #1693, #1694, and #1695, in Table 2).

On the ROCOv2 benchmark our system surpasses single-encoder and concept-agnostic baselines on every shared-task metric, BERTScore, ROUGE-1, BLEURT, ALIGN, and UMLS-F1, demonstrating simultaneous gains in linguistic relevance and clinical factuality. Among the post-processing strategies, BioBERT-centric reranking achieves the best harmonic mean of relevance and factuality, whereas BioMedCLIP-based scoring offers the highest image-text alignment, highlighting a trade-off that can be tuned to downstream needs.

These results validate two key insights: (1) heterogeneous visual encoders supply complementary features that improve descriptive richness, and (2) explicit concept supervision curbs hallucinations and improves diagnostic grounding. The proposed pipeline establishes a strong baseline for upcoming CLEF-Cap tasks and paves the way for future work on longitudinal captioning, device detection, and lightweight on-device deployment in clinical settings.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00360176).

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI GPT-4o in order to: refine writing style, reorganize paragraph structure, and assist with technical language formulation. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. R. Beddiar, M. Oussalah, T. Seppänen, Automatic captioning for medical imaging (mic): a rapid review of literature, Artificial intelligence review 56 (2023) 4019–4076.

[2] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, Y. Gu, A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations, Expert Systems with Applications 242 (2024) 122807.

[3] M. Limbu, D. Banerjee, Medblip: Fine-tuning blip for medical image captioning, arXiv preprint arXiv:2505.14726 (2025).

[4] H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey, IEEE Transactions on Biomedical Engineering 69 (2021) 1173–1185.

[5] S. Umirzakova, S. Ahmad, L. U. Khan, T. Whangbo, Medical image super-resolution for smart healthcare applications: A comprehensive survey, Information Fusion 103 (2024) 102075.

[6] F. Pérez-García, H. Sharma, S. Bond-Taylor, K. Bouzid, V. Salvatelli, M. Ilse, S. Bannur, D. C. Castro, A. Schwaighofer, M. P. Lungren, et al., Exploring scalable medical image encoders beyond text supervision, Nature Machine Intelligence (2025) 1–12.

[7] Z. Lu, H. Li, N. A. Parikh, J. R. Dillman, L. He, Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining, IEEE Transactions on Neural Networks and Learning Systems (2025).
[8] P. Kaliosis, J. Pavlopoulos, F. Charalampakos, G. Moschovis, I. Androutsopoulos, A data-driven guided decoding mechanism for diagnostic captioning, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7450–7466. URL: https://aclanthology.org/2024.findings-acl.444/. doi:10.18653/v1/2024.findings-acl.444.
[9] L. Chen, X. Yang, G. Jeon, M. Anisetti, K. Liu, A trusted medical image super-resolution method based on feedback adaptive weighted dense network, Artificial Intelligence in Medicine 106 (2020) 101857.
[10] A. Telea, An image inpainting technique based on the fast marching method, Journal of graphics tools 9 (2004) 23–34.
[11] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).
[12] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, arXiv preprint arXiv:2303.00915 (2023).
[13] Z. Zhang, B. Wang, W. Liang, Y. Li, X. Guo, G. Wang, S. Li, G. Wang, Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 1731–1735.
[14] 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.
[15] ContactDoctor, ContactDoctor-Bio-Medical: A High-Performance Biomedical Language Model, https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024. Accessed: 2025-06-16.
[16] D. M. Chan, A. Myers, S. Vijayanarasimhan, D. A. Ross, J. Canny, Ic3: Image captioning by committee consensus, arXiv preprint arXiv:2302.01328 (2023).
[17] T. Sellam, D. Das, A. P. Parikh, Bleurt: Learning robust metrics for text generation, arXiv preprint arXiv:2004.04696 (2020).
[18] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, S. E. A. Ouatik, An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings, Expert Systems with Applications 167 (2021) 114152.
[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[20] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: A. Rumshisky, K. Roberts, S. Bethard, T. Naumann (Eds.), Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://aclanthology.org/W19-1909/. doi:10.18653/v1/W19-1909.
[21] Q. Han, S. Tian, J. Zhang, A pubmedbert-based classifier with data augmentation strategy for detecting medication mentions in tweets, arXiv preprint arXiv:2112.02998 (2021).
[22] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.
[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation

with BERT, in: ICLR, 2020.
[26] E. C. Seyhan, S. N. Sokucu, G. Gunluoglu, N. S. Veske, S. Altin, Primary pulmonary synovial sarcoma: a very rare presentation, Case Reports in Pulmonology 2014 (2014) 537618–537618.
[27] O. N. Al Mulhim, Huge thoracic aortic aneurysm presenting with jaundice: A case report, Vascular Health and Risk Management (2022) 1–4.
[28] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[29] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[30] E. Hirsch, G. Dawidowicz, A. Tal, Medrat: Unpaired medical report generation via auxiliary tasks, in: European Conference on Computer Vision, Springer, 2024, pp. 18–35.
[31] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[32] 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.
[33] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.
[34] 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.
[35] 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.
[36] 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).
[37] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, arXiv preprint arXiv:1511.06709 (2015).

Appendix A. Summarization-based Refinement Results

While summarization-based refinement using GPT models provides a promising approach for aggregating multiple candidate captions into a single concise output, as summarized in Table 4, our experiments reveal that this strategy underperforms compared to reranking-based methods in terms of factual alignment and clinical adequacy.

The best summarization method — CoT-based refinement (#1938) — achieves a BERTScore of 0.5705, BLEURT of 0.3197, ALIGN of 0.0843, and UMLS F1 of 0.1236. In comparison, the BioMedCLIP-based reranking method (Table 3), #1900) outperforms it across all key metrics, achieving a higher BERTScore

Table 4
Quantitative comparison of GPT-based summarization methods for medical caption refinement.

Submission ID	Method	Image-text SIM	BERTScore (Recall)	ROUGE-1	BLEURT	ALIGN	UMLS F1
#1938	CoT	0.8628	0.5705	0.2032	0.3197	0.0843	0.1236
#1943	Prompt guided	0.8638	0.5723	0.2016	0.3176	0.0795	0.1242

(0.5823 vs. 0.5705, +0.0118), ROUGE (0.2440 vs. 0.2032, +0.0408), ALIGN (0.1231 vs. 0.0843, +0.0388), and UMLS F1 (0.1524 vs. 0.1236, +0.0288). Although BLEURT is marginally higher in the CoT setting (0.3197 vs. 0.3173), the gains in factuality and alignment metrics clearly favor reranking.

This discrepancy stems from the nature of the two approaches: while reranking explicitly evaluates and selects candidates based on semantic similarity with image content, summarization methods rely on generative synthesis, which may abstract away or omit critical clinical entities during compression. The notable drop in ALIGN and UMLS F1 supports this interpretation, indicating that summarization may weaken grounding by generalizing away from medically specific content.

Reranking methods — particularly those leveraging visual-textual alignment like BioMedCLIP — are more effective at preserving semantic fidelity and factual grounding in medical image captioning. Summarization, although useful for improving fluency, should be applied with caution in high-precision clinical settings.