ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Bachelor Thesis
in
Computer Science

Exploring Post-Training Techniques for Diagnostic Captioning

Ippokratis Pantelidis

Supervisors: Ion Androutsopoulos

John Pavlopoulos

September 2025

Ippokratis Pantelidis
Exploring Post-Training Techniques for Diagnostic Captioning
September 2025
Supervisors: Ion Androutsopoulos, John Pavlopoulos

Athens University of Economics and Business
School of Information Sciences and Technology
Department of Informatics
Information Processing Laboratory, Natural Language Processing Group
Athens, Greece

Abstract

Image Captioning is a research area at the intersection of Computer Vision (CV) and Natural Language Processing (NLP), focusing on the automatic generation of descriptive text for images. In the medical field, this task becomes especially important when applied to diagnostic images such as radiographs. Known as Diagnostic Captioning (DC), the goal is to generate clinically meaningful text that reflects a patient's condition based on visual input. This thesis investigates how general-purpose vision-language models can be adapted to meet the specific demands of this high-stakes application. Three post-training strategies are explored: Supervised Fine-Tuning (SFT), Reinforcement Learning (RL), and Test-Time Scaling (TTS). These approaches are applied to state-of-the-art models and evaluated using a dedicated benchmark from the ImageCLEFmedical 2025 challenge. The systems developed in this thesis are assessed both quantitatively, using relevance and factuality metrics, and qualitatively, through examples that demonstrate model behavior under different conditions. The results highlight the strengths and limitations of each approach, and also show that these methods can be effectively combined to leverage the advantages of each technique. Overall, the work contributes to ongoing efforts in making AI-assisted diagnosis more reliable in real-world medical settings.

Περίληψη

Η αυτόματη περιγραφή εικόνων αποτελεί έναν ερευνητικό τομέα που βρίσκεται στην τομή της Υπολογιστικής Όρασης (Computer Vision, CV) και της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing, NLP). Στον ιατρικό τομέα, η τεχνολογία αυτή αποκτά ιδιαίτερη σημασία όταν εφαρμόζεται σε διαγνωστικές εικόνες, όπως είναι οι ακτινογραφίες. Το πεδίο αυτό αναφέρεται συχνά ως Διαγνωστική Παραγωγή Λεζάντας (Diagnostic Captioning, DC), με βασικό στόχο την αυτόματη δημιουργία κειμένων που αποτυπώνουν με κλινική ακρίβεια την κατάσταση του ασθενούς βάσει οπτικής πληροφορίας. Η παρούσα πτυχιακή εργασία εξετάζει τρόπους προσαρμογής γενικών πολυτροπικών μοντέλων όρασης-γλώσσας στις απαιτήσεις της διαγνωστικής περιγραφής, εστιάζοντας σε τρεις στρατηγικές μεταεκπαίδευσης: το Supervised Fine-Tuning (SFT), την Ενισχυτική Μάθηση (Reinforcement Learning, RL), και το Test-Time Scaling (TTS). Οι μέθοδοι αυτές εφαρμόζονται σε μοντέλα αιχμής και αξιολογούνται μέσω του συνόλου δεδομένων του διαγωνισμού ImageCLEFmedical 2025. Τα αναπτυχθέντα συστήματα αξιολογούνται τόσο ποσοτικά, με χρήση μετρικών που καλύπτουν τη συνάφεια και την κλινική ακρίβεια, όσο και ποιοτικά, μέσω παραδειγμάτων που αναδεικνύουν τη συμπεριφορά των μοντέλων υπό διαφορετικές συνθήκες. Τα αποτελέσματα αναδεικνύουν τα πλεονεκτήματα και τους περιορισμούς κάθε μεθόδου και δείχνουν ότι ο συνδυασμός τους μπορεί να αξιοποιηθεί αποτελεσματικά, ενισχύοντας τα πλεονεκτήματα της καθε μίας. Συνολικά, η εργασία αυτή συμβάλλει στην πρόοδο της τεχνητής νοημοσύνης στον τομέα της ιατρικής και προωθεί την ανάπτυξη συστημάτων που μπορούν να υποστηρίξουν αποτελεσματικά τη διαγνωστική διαδικασία.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisors, Ion Androutsopoulos and John Pavlopoulos, for the opportunity to work under their guidance. Their valuable feedback, continuous support, and encouragement were essential throughout this thesis and greatly influenced my academic growth. I am also grateful to PhD students Giorgos Moschovis, Foivos Charalampakos, and Panagiotis Kaliosis for their time and guidance. Their advice and discussions were extremely helpful in both the theoretical and practical aspects of this work. Special thanks go to MSc student Anna Chatzipapadopoulou and Marina Samprovalaki, Research Assistant at the AUEB NLP Group, for their consistent collaboration and insightful conversations throughout the thesis. Their support played an important role in its development. Finally, I want to thank my family and close friends for always being by my side. Their encouragement and belief in me were a constant source of strength and motivation.

Contents

1 Introduction

In recent years, advances in medical imaging technologies have significantly enhanced the ability of healthcare systems to detect, monitor, and diagnose a wide range of conditions. Modern imaging modalities such as X-rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) generate an ever-growing volume of high-resolution diagnostic data. However, this rapid expansion has placed increasing pressure on radiologists, who are required to manually interpret large numbers of complex medical images. The resulting workload can lead to delays in reporting, increased fatigue, and a higher likelihood of diagnostic errors, particularly in time-sensitive clinical environments [Kas+23]. Diagnostic Captioning (DC) is a specialized area of image captioning focused on generating clinically relevant textual reports or summaries from medical images, such as X-rays or CT scans [Pav+21]. Unlike text-to-text models, multimodal DC systems combine visual processing with natural language generation by extracting meaningful features from images and translating them into clinically informative text. These systems are designed to assist radiologists by providing preliminary draft reports, suggesting areas of interest to guide their attention, or drawing attention to image regions that may warrant further review. Rather than replacing expert interpretation, DC aims to reduce reporting time, enhance diagnostic consistency, and ultimately contribute to improved patient outcomes.

This thesis explores the use of post-training techniques in the context of DC. Advances in Deep Learning (DL) have driven remarkable progress in both Computer Vision (CV) and Natural Language Processing (NLP), leading to the development of powerful multimodal models capable of interpreting images and generating coherent text. These models are typically trained in two stages: an initial pre-training phase, where they are exposed to large-scale collections of image-text pairs to learn general visual-linguistic representations; and a subsequent post-training phase, which adapts these representations to specific domains or tasks [SD25]. Post-training encompasses a range of methods designed to refine model performance, enhance reasoning capabilities, or better align outputs with human expectations. According to a recent survey by Wei et al. [Wei+23], these methods can be broadly categorized into three classes: Supervised Fine-Tuning (SFT) on specialized datasets, Reinforcement Learning (RL) approaches that optimize task-specific objectives, and Test-Time Scaling (TTS) strategies, which improve predictions at inference time without modifying model parameters. Such techniques are particularly relevant to Diagnostic Captioning (DC), where factual accuracy (the correctness of the information presented), clinical precision (the detailed and accurate description of medical findings), and domain sensitivity (the model's ability to respect medical terminology and context)

are critical. These metrics ensure that generated captions are reliable and useful for clinical decision-making, minimizing the risk of misinterpretation or incorrect diagnoses that could adversely affect patient care. By applying post-training methods to pretrained Vision-Language Models (VLMs), this thesis aims to bridge the gap between general-purpose captioning systems and the specialized demands of the medical domain.

Part of this thesis focuses on the participation of the AUEB NLP Group in the ImageCLEFmedical Caption Task 2025 [Dam+25], organized as part of the broader ImageCLEF 2025 campaign [Ion+25]. The Caption Task consists of three primary sub-tasks: Concept Detection, which involves predicting relevant medical concepts associated with an image; Caption Prediction, which aims to generate coherent and clinically informative descriptions; and the Explainability Task, which focuses on providing human-interpretable justifications for the model's predictions. The author's main responsibility was the development of systems for the Caption Prediction sub-task, which also forms the core topic of this thesis. Building on the group's strong track record in previous editions of the competition [Cha+21; Cha+22; Kal+23; Sam+24], this year the AUEB NLP Group achieved 1st place in Concept Detection, 5th place in Caption Prediction, and 1st place in the Explainability Task among 9, 8, and 2 participating research groups, respectively [Cha+25].

1.1 Thesis Structure

Chapter 2: Background and Related Work

Chapter 2 reviews foundational concepts in vision-language models and presents prior work on medical image captioning.

Chapter 3: Implemented Methods and Systems

Chapter 3 describes the models and post-training techniques implemented in this thesis, including fine-tuning, reinforcement learning, and test-time scaling.

Chapter 4: Data

Chapter 4 presents the dataset used in this thesis and provides an exploratory analysis of the captioning data relevant to the ImageCLEFmedical 2025 challenge.

Chapter 5: Experiments and Results

Chapter 5 reports both qualitative and quantitative results of the developed systems, along with insights from participation in the ImageCLEFmedical 2025 competition.

Chapter 6: Conclusions and Future Work

Chapter 6 summarizes the main findings and outlines potential directions for future research.

2 Background and Related Work

This chapter provides an overview of the key concepts and research areas that underpin this thesis. It first introduces the foundational training paradigms of large-scale multimodal systems, including pre-training and post-training. It then reviews related work on generic image captioning and explores the distinct challenges posed by diagnostic captioning in the medical domain.

The rapid evolution of DL has enabled the emergence of powerful multimodal systems capable of jointly processing visual and textual information. These models, commonly referred to as VLMs, integrate advances from both CV and NLP to support tasks such as image captioning, retrieval, and reasoning [BAM17]. Their development is driven by the broader progress in Large Language Models (LLMs), which have demonstrated strong generalization capabilities when trained on vast corpora of text [Bro+20]. The learning process for these systems typically begins with a large-scale pre-training phase designed to capture general visual-linguistic representations [Zho+20]. However, such general-purpose training often falls short in high-stakes domains, requiring a subsequent post-training phase to adapt models to specific applications [Wei+23]. One of the most prominent use cases for VLMs is image captioning [Hos+18], where models generate natural language descriptions for visual content. While generic captioning models can produce fluent and semantically relevant output, transferring these capabilities to diagnostic captioning involves additional requirements, such as clinical accuracy, appropriate use of terminology, and interpretability [Pav+21]. Bridging this gap remains a central challenge for research in medical Artificial Intelligence (AI).

2.1 Pre-Training

This section begins by examining pre-training techniques for textual data, which laid the groundwork for the development of VLMs, a subset of multimodal models specialized in processing images and text. In NLP, pre-training has become a foundational paradigm, largely driven by advances in Self-Supervised Learning (SSL) applied to LLMs. The central idea is to expose models to large volumes of unlabeled text, enabling them to learn general-purpose language representations that can later be adapted to specific downstream tasks via fine-tuning or other post-training methods (Section 2.2).

Several milestone models exemplify this shift, including BERT [Dev+19], GPT-2/3 [Bro+20], and T5 [Raf+19]. Each introduced novel self-supervised objectives and architectures that significantly advanced the state of the art across a wide range of NLP tasks. To illustrate these pre-training dynamics more concretely, we focus on BERT, one of the most influential early models. BERT is trained using two self-supervised objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). As shown in Figure 2.1, MLM involves randomly masking a subset of input tokens and training the model to predict the original tokens based on their surrounding context. This objective encourages the model to learn bidirectional contextual representations at the word level. In contrast, NSP is designed to help the model capture discourse-level coherence by predicting whether two input sentences appear consecutively in the source text.

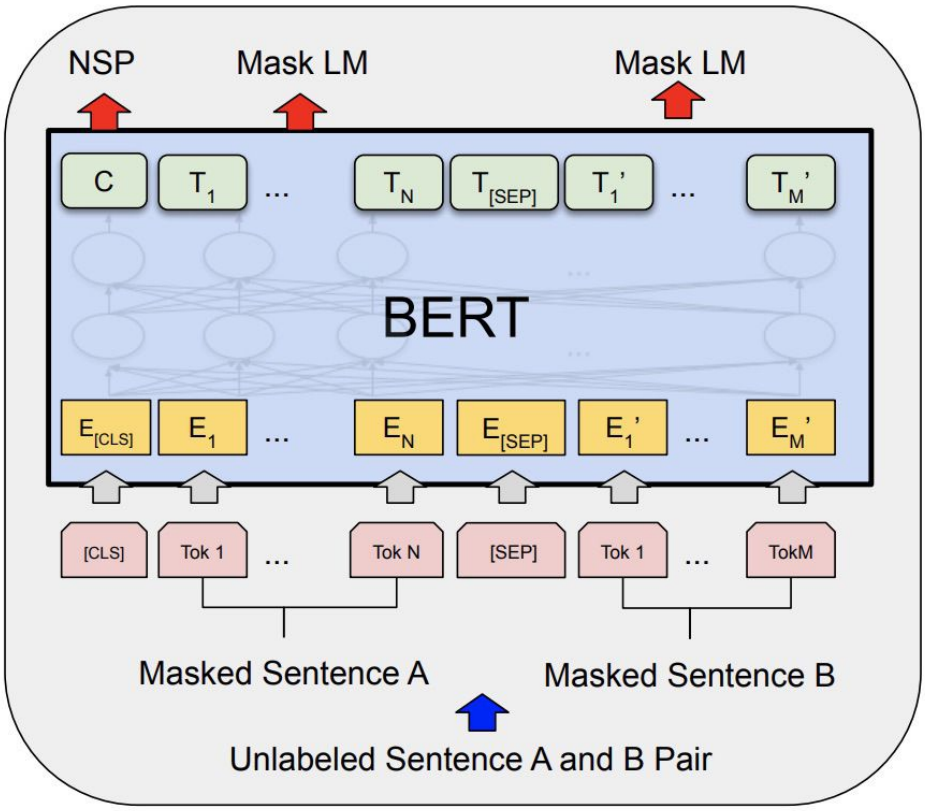


Fig. 2.1: Illustration of the Masked Language Modeling (MLM) objective used in the pre-training of BERT. Tokens are randomly masked and predicted from surrounding context. Figure taken from [Dev+19].

The success of BERT [Dev+19] and similar models in NLP inspired a new class of architectures capable of handling both visual and textual modalities, known as VLMs. These models are trained on paired image-text data to enable tasks such as image captioning, cross-modal retrieval, and visual question answering. Unlike language-only models, VLMs incorporate visual inputs using either dual-encoder architectures, where separate encoders are used for each modality, or unified transformer-based models that jointly process multimodal inputs.

Pre-training visual-language models (VLMs) generally depends on large-scale datasets comprising millions of image-caption pairs. Some of the most commonly used datasets include Conceptual Captions [Sha+18], COCO [Lin+14], and LAION [Sch+22]. Conceptual Captions is automatically constructed from web alt-text and contains around 3.3 million pairs; while large, it is weakly supervised and prone to noise. COCO, in contrast, is a smaller but higher-quality dataset consisting of over 330,000 images with five manually written captions each, commonly used for fine-tuning and evaluation. LAION offers web-scale coverage, containing hundreds of millions of image-text pairs collected and filtered automatically using CLIP embeddings, which are discussed below. Though noisier, its massive scale makes it particularly suitable for contrastive pre-training. To learn aligned multimodal representations from such data, VLMs are typically optimized using a combination of self-supervised objectives. Contrastive Learning (CL), as in Contrastive Language–Image Pre-training (CLIP) [Rad+21], learns a shared embedding space by pulling matched image–text pairs closer together while pushing mismatched ones apart, typically across all pairs in a batch. In contrast, Image-Text Matching (ITM) is formulated as a binary classification task: given a single image-caption pair, the model predicts whether the caption is semantically aligned with (i.e., describes) the image, rather than optimizing over multiple pairs as in supervised contrastive learning. While CL encourages global alignment in the embedding space, ITM enforces pairwise discrimination, providing complementary supervision for fine-grained alignment. Masked Image Modeling (MIM) trains the model to reconstruct masked patches in the input image, improving spatial understanding. Finally, Multimodal Masked Language Modeling (MMLM) extends the BERT-style MLM objective to the multimodal setting, requiring the model to fill in masked text tokens using both visual and linguistic context. Together, these objectives enable VLMs to learn rich, transferable representations suitable for a wide range of downstream tasks.

A prominent realization of the contrastive pre-training paradigm is CLIP [Rad+21], already briefly mentioned above. More concretely, CLIP introduces a scalable dual-encoder framework where visual and textual inputs are processed separately: a Vision Transformer (ViT) or ResNet [He+16] encodes the image, and a Transformer-based encoder handles the corresponding text. The model is trained to project both modalities into a shared embedding space, such that matching image-caption pairs are close, while mismatched pairs are pushed apart. This alignment is achieved using a contrastive loss computed over all $N \times N$ pairwise similarities in a batch of N image-text examples. A symmetric cross-entropy objective is applied in both directions (image-to-text and text-to-image), encouraging strong multimodal association without the need for explicit labels. Figure 2.2 illustrates the architecture and training setup of CLIP. By using separate encoders and a contrastive loss across the entire batch, CLIP achieves efficient and scalable learning across noisy web data. It was trained on 400 million image-text pairs collected from the Internet, allowing it to generalize in a zero-shot manner to a wide range of tasks without task-specific fine-tuning. This approach has since influenced a broad family of vision-language models aiming for similar flexibility and scale.

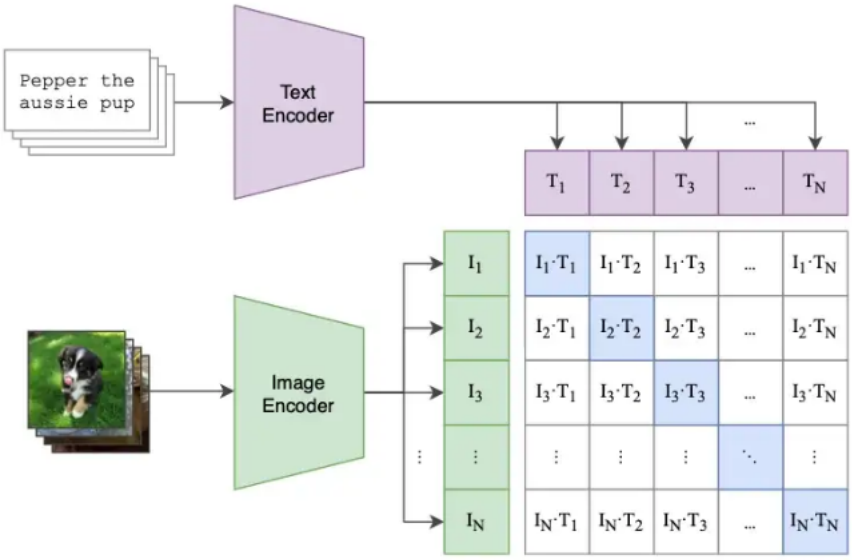


Fig. 2.2: Contrastive pre-training in CLIP [Rad+21]. A text encoder and an image encoder independently map their inputs into a joint embedding space. During training, similarity is maximized for aligned image-text pairs and minimized for all others using a contrastive objective. Figure taken from [Rad+21].

Recent models such as BLIP [Li+22], Flamingo [Ala+22], and InstructBLIP [Dai+23], which will be discussed in detail later in this thesis, extend the pre-training paradigm beyond dual-encoder designs by adopting more integrated transformer-based architectures. These systems combine multiple learning signals during training and exhibit strong performance on open-ended generation and reasoning tasks, even under minimal supervision.

In summary, pre-training has become a cornerstone of modern vision-language modeling, enabling the development of systems that generalize across a wide range of tasks with minimal supervision. It relies on large-scale datasets that are typically collected automatically and require little or no human annotation, offering scalability and flexibility at a global scale. However, pre-training is also a resource-intensive process, often performed only once due to its high computational cost and environmental impact. The resulting models are reused across many downstream applications, making the design of effective pre-training strategies a critical foundation for later stages of model adaptation. For this reason, understanding the role and implications of pre-training is essential, even in work such as this thesis that focuses primarily on post-training refinement.

2.2 Post-Training

While pre-training endows language and vision-language models with broad general-purpose capabilities, it often proves insufficient for high-stakes or domain-specific applications, such as medical image interpretation or diagnostic reporting. Post-training refers to

the set of methods applied after the pre-training phase to further adapt models to specific tasks, align their behavior with human intent, or enhance their reasoning and factual accuracy. According to the taxonomy proposed by Wei et al. [Wei+23], post-training techniques can be broadly grouped into three categories: Supervised Fine-Tuning (SFT), Reinforcement Learning (RL), and Test-Time Scaling (TTS) strategies. This classification is illustrated in Figure 2.3, which shows a four-layer taxonomy: the innermost layer highlights the three main categories (SFT, RL, and TTS), the next layer outlines their sub-dimensions, the third layer lists representative algorithms, and the outermost layer maps well-known LLMs and VLMs to the strategies they use.

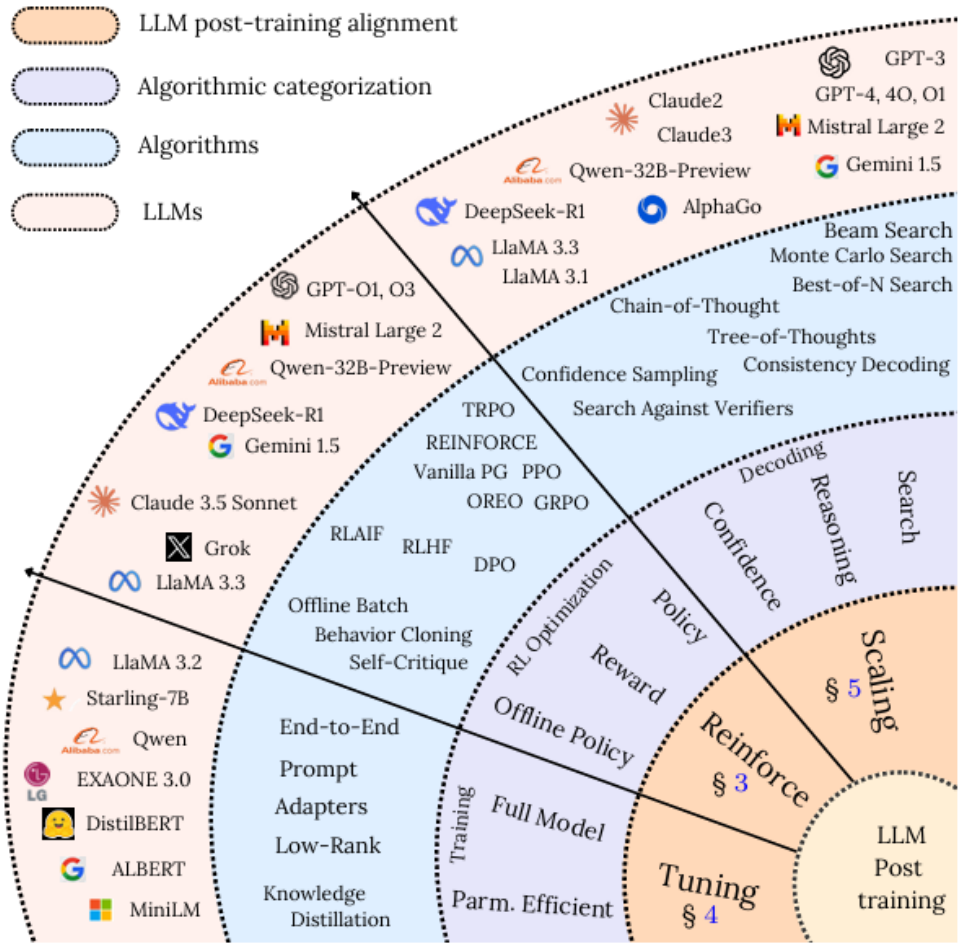


Fig. 2.3: Taxonomy of post-training techniques in LLMs and VLMs, including Supervised Fine-Tuning (SFT), Reinforcement Learning (RL), and Test-Time Scaling (TTS) strategies. Figure taken from [Wei+23].

2.2.1 Supervised Fine-Tuning

SFT is the most widely used post-training technique and serves as the primary mechanism for adapting pre-trained models to specific tasks or domains. It involves training the model on labeled input-output pairs using standard supervised learning objectives, most commonly the cross-entropy loss. Unlike full-scale model training, SFT involves only

modest adjustments to the model. It typically uses fewer training epochs (e.g., 3-5 for BERT-style models), smaller learning rates, and often updates only a subset of parameters. This careful, constrained training prevents catastrophic forgetting while allowing the model to internalize task-specific patterns, domain terminology, stylistic conventions, and structural regularities in the target domain. In the context of DC, SFT plays a critical role by exposing the model to clinically annotated image-text pairs, allowing it to learn accurate medical phrasing, domain-specific vocabulary, and clinically relevant details.

The success of SFT largely depends on the availability and quality of labeled data [Wan+23b]. Rich, diverse annotations can significantly improve model performance and generalization, while limited or noisy supervision may lead to overfitting or bias. Regularization techniques, data augmentation, and transfer learning strategies are often employed to mitigate these risks. Despite its simplicity, SFT remains a powerful and flexible method for aligning general-purpose models with the demands of specialized tasks. To better illustrate the concept of SFT, Figure 2.4 provides a high-level view of the process. An LLM is initially pre-trained on a massive web-scale corpus to learn general-purpose representations. This base model is then adapted to a downstream task by fine-tuning it on a smaller, domain-specific dataset. In this case, the supervised fine-tuning stage incorporates task-specific knowledge from a private or curated source, such as clinical annotations or task-specific text, allowing the model to internalize relevant patterns, terminology, and output conventions. The result is a model that preserves its broad linguistic competence while gaining specialization for the target application.

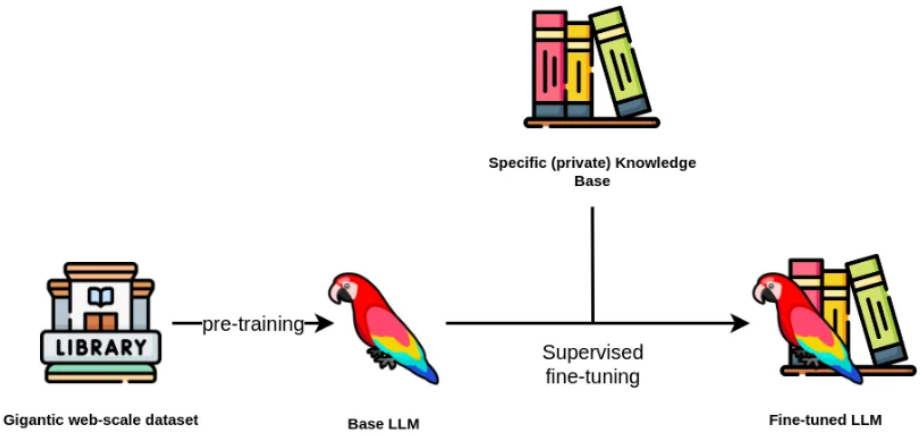


Fig. 2.4: Conceptual overview of Supervised Fine-Tuning (SFT). A base language model pre-trained on web-scale data is adapted using a smaller, domain-specific dataset to specialize for downstream tasks. Figure adapted from Tomaž Bratanič and Kumar Harsh, illustrating the general SFT workflow. Knowledge Graphs and LLMs: Fine-Tuning vs. Retrieval-Augmented Generation, Neo4j Blog, September 11, 2024. `https://neo4j.com/blog/developer/fine-tuning-vs-rag/`.

2.2.2 Reinforcement Learning

Reinforcement Learning (RL) is a learning paradigm in which an agent interacts with an environment and learns to make decisions by maximizing a reward signal. Unlike supervised learning, where the model is trained on labeled input-output pairs, RL focuses on learning optimal actions through trial and error, guided by feedback about the quality of its predictions. In the context of language and vision-language models, RL is particularly useful for optimizing objectives that are non-differentiable or difficult to encode directly, such as factual consistency, clinical accuracy, or human preferences [SB18].

Reinforcement Learning with Human Feedback (RLHF) [Ouy+22] has become a widely used method, where human annotators rank model outputs to train a reward model. This reward model serves as a stand-in for human judgment and is then used to steer policy optimization, typically through policy gradient methods that adjust the model's parameters to increase the likelihood of producing preferred responses. RLHF has played a key role in aligning LLMs with human values and task expectations, especially when the evaluation criteria are hard to define or measure directly. It has also shown effectiveness in reducing hallucinations by steering models away from factually incorrect or misleading outputs.

In image captioning, and in particular in the domain of DC, reinforcement learning (RL) can be used to directly optimize task-specific reward functions. Standard training with cross-entropy loss often fails to align with clinical correctness or informativeness [Gao+19]. By contrast, RL enables models to optimize evaluation metrics more directly, such as BERTScore [Zha+20], as shown by the winning system at ImageCLEFmedical 2023 [NDK23], or more domain-oriented metrics like UMLS Concept F1. In some cases, a weighted combination of multiple objectives is used to better align training with clinical goals, as it will be discussed in Section 3.3. This ability to define flexible, domain-sensitive rewards makes RL a powerful tool for fine-tuning captioning models in clinical settings.

Beyond RLHF and reward modeling based on external metrics, RL has also been applied more directly to image captioning through actor–critic methods. In these setups, a policy network incrementally generates a caption by predicting the next word given the current state, while a value network estimates the expected cumulative reward from that state onward. This approach enables lookahead inference, where the model evaluates the potential quality of future tokens and uses this signal to improve generation. Rewards are typically defined using semantic similarity to reference captions, embedding-based alignment, or task-specific heuristics. Figure 2.5 illustrates this process in detail. The current state is represented by the partially generated caption ("a cat is") alongside the input image. The policy network takes this state and produces a distribution over possible next actions (candidate words such as "lying", "sitting", "holding", etc.). At the same time, the value network predicts the expected reward associated with continuing from that state.

By combining these two signals, the system can not only choose the most likely next word but also anticipate how that choice will influence the quality of the complete caption. For instance, in the example shown, the model considers options like "lying" or "sitting" but selects "holding", leading toward the caption "a cat is holding a baseball bat". The key point is that the actor–critic framework allows the captioning system to optimize word-by-word generation with respect to a long-term reward signal, rather than relying solely on local word probabilities.

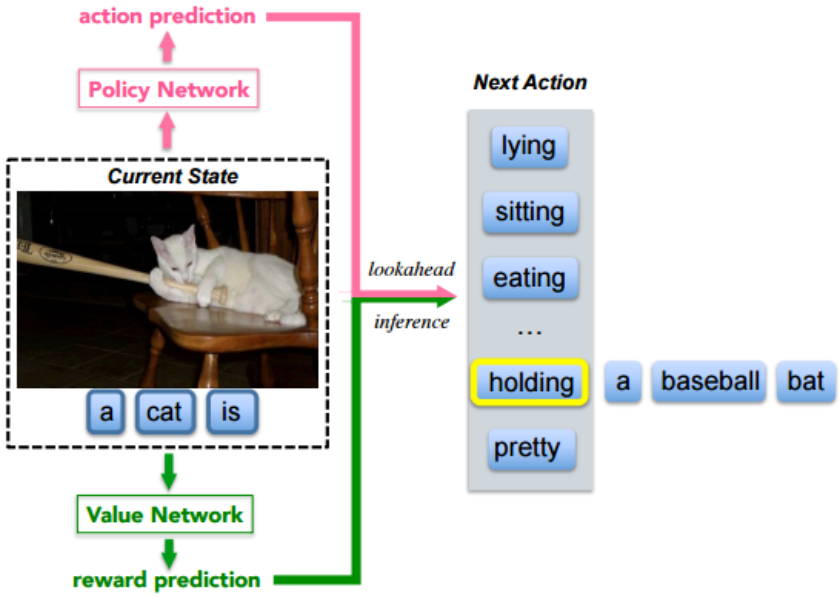


Fig. 2.5: Actor–Critic framework for image captioning. At each time step, the policy network generates the next word in the caption, while the value network predicts the expected future reward based on that partial sequence. The two networks are trained jointly to produce captions that maximize a task-specific reward signal. Figure taken from [Ren+17a].

However, RL methods come with notable challenges. They are typically more computationally expensive than supervised fine-tuning, require extensive sampling, and are highly sensitive to the reward design. Poorly calibrated rewards may lead to degenerate or unstable behavior, especially in safety-critical domains like healthcare.

2.2.3 Test-Time Scaling

Test-Time Scaling (TTS) refers to methods applied during inference to improve model performance without modifying its parameters. These techniques offer a lightweight and flexible alternative to SFT or RL, making them especially valuable when labeled data is limited, computational resources are constrained, or model updates are impractical.

Prompting can be viewed as a form of TTS, where carefully designed input instructions are used to steer model behavior. While prompting is more often discussed in the context

of language modeling, it can also be understood as part of the broader family of TTS methods. Even minor prompt modifications, such as rephrasing a question or adding cues like "Let's think step by step," can significantly improve model outputs [Koj+22]. This effect is especially pronounced in LLMs, where zero-shot or few-shot prompts can elicit more structured or accurate responses. A prominent example is Chain-of-Thought (CoT) prompting [Wei+22b], which encourages intermediate reasoning by structuring prompts to model multi-step inference processes.

Beyond text-only tasks, TTS techniques have also proven effective in VLMs and image captioning. Common strategies include:

- **Caption reranking:** Generating multiple caption candidates and selecting the best one based on external scoring models or domain-specific heuristics [Wei+23], a concept that also underpins one of the methods discussed in Section 3.4.

- **Iterative decoding:** A method in which the model first generates an initial caption and then refines it over multiple decoding passes, gradually improving coherence and factual accuracy [Wei+23].

- **Retrieval-augmented generation (RAG):** At inference time, retrieving relevant context, such as image-caption pairs, from a datastore to guide caption generation [REM23].

- **Controllable generation:** Guiding the model's output during inference by modifying prompts or inserting control tokens to enforce stylistic, structural, or domain-specific constraints [Kes+19].

In the context of DC, TTS methods are particularly valuable when labeled supervision is limited or domain shift is expected at deployment. For instance, retrieving captions from visually similar medical images can provide soft guidance, serving as contextual references to improve output quality without modifying the underlying model. A representative application of this concept is the Synthesizer module, introduced in [Sam+24]. The Synthesizer operates in two stages. It first receives a draft caption from a base model and then retrieves one or more image–caption pairs from the dataset based on visual similarity to the test image. However, experiments show that using a single retrieved caption yields the best results. It subsequently generates a refined caption that is informed both by the input image and the retrieved reference. This process can be understood as a form of RAG-style TTS, where retrieval is used not to augment prompts directly, but to guide the generation toward captions that are more consistent with visually similar examples. In doing so, the Synthesizer [Sam+24] facilitates inference-time adaptation to local data distributions without requiring any changes to the model parameters.

Another Test-Time Scaling (TTS) approach proposed by the AUEB NLP Group is a concept-driven decoding mechanism introduced by Kaliosis et al. [Kal+24]. This method enhances DC by guiding the generation process using predicted medical tags associated with the input image. These tags, which represent key medical concepts, are used during inference to influence the beam search decoding process, encouraging the model to include semantically relevant content in the output. The technique, named DMMCS (Distance from Median Maximum Concept Similarity) [Kal+24], imposes a soft penalty on token choices that deviate from the expected semantic association between tags and caption content, based on statistics learned from the training set. The method has shown improvements across multiple architectures and datasets, even when the tags are noisy or automatically predicted, as demonstrated by its extension to handwritten text recognition [KP25]. More details on how these tags are acquired can be found in [Cha25] and in our group's participation in the ImageCLEFmedical 2025 challenge [Cha+25], specifically in the Concept Detection task. DMMCS represents a compelling example of controllable generation applied at test time, offering improved clinical accuracy and domain specificity without requiring any modification to the model parameters.

While TTS methods are typically lightweight and efficient, their effectiveness often depends on the quality of the prompt, retrieval corpus, or reranking strategy. Still, they offer a promising avenue for real-world deployment, particularly in clinical scenarios where continual fine-tuning may be impractical or risky.

2.3 Generic Image Captioning

Image captioning is the task of generating descriptive natural language sentences that summarize the visual content of a given image. As a classic example of a multimodal challenge, it requires integrating visual perception with linguistic reasoning. This capability is broadly useful in real-world applications such as assisting visually impaired users in understanding visual content, improving accessibility in web and mobile platforms, enhancing image organization and search through automatic metadata generation, and enabling content moderation and media summarization.

Formally, given an image $\mathcal{I}$, the objective is to produce a caption $\hat{y} = (y_1, y_2, \ldots, y_T)$, where each y_t represents a word token and T is the length of the generated sequence. The caption should accurately reflect the salient entities, actions, and attributes present in the scene. Figure 2.6 shows examples of three successful captions generated by modern captioning systems across a range of generic, everyday life images drawn from diverse visual contexts.

[1] `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning`

(a) A dog running around in a field with a soccer ball

(b) A red car is driving down the street

(c) A cup of coffee and a pastry on a plate

Fig. 2.6: Examples of image–caption pairs created by the ViT-GPT2 Image Captioning model[1].

Historically, early approaches to image captioning followed an encoder-decoder framework, inspired by sequence-to-sequence models used in machine translation. A seminal work in this direction is the "Show and Tell" model [Vin+15], where a Convolutional Neural Network (CNN) such as Inception [Sze+15] or ResNet [He+16] serves as the image encoder, and a Recurrent Neural Network (RNN) [RHW86], typically a Long Short-Term Memory (LSTM) network [HS97], generates the caption sequentially. As illustrated in Figure 2.7, the model is trained end to end to maximize the likelihood of the target caption given the image embedding. This architecture established a foundational pipeline and demonstrated the feasibility of learning grounded descriptions from data.

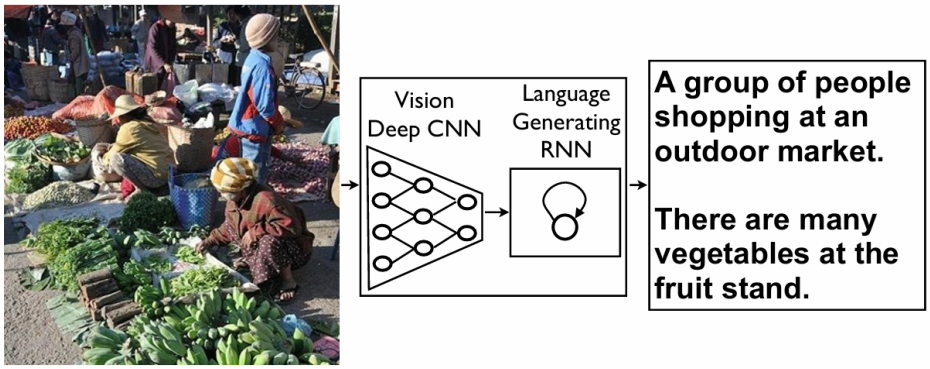


Fig. 2.7: Architecture of the Show and Tell model [Vin+15], where an image is encoded using a CNN and decoded into a caption using an LSTM network. Figure taken from [Vin+15].

However, basic CNN–RNN models had limitations in spatial reasoning and lacked the ability to focus on specific regions of an image. This led to the development of attention-based methods, most notably the "Show, Attend and Tell" model [Xu+15], which introduced a soft visual attention mechanism. Rather than encoding the image into a single global feature vector, this model computes attention weights over convolutional feature maps at each time step of the caption generation process. This enables the decoder to dynamically focus on different spatial regions depending on the word being generated, significantly improving both the descriptive richness and interpretability of the captions. As illustrated in Figure 2.8, the model begins by extracting convolutional features from the input image. These features are then fed into an attention-equipped LSTM decoder, which enables the model to focus on specific regions of the image at each time step. As the model

generates words sequentially, attention visualizations highlight the relevant parts of the image corresponding to each word.

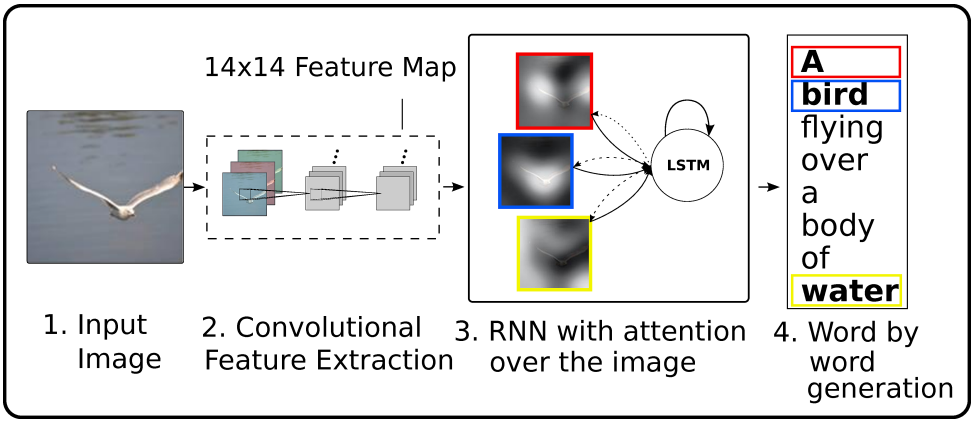


Fig. 2.8: Overview of the Show, Attend and Tell model [Xu+15], which introduces soft visual attention. The model dynamically attends to different spatial regions of the image while generating each word in the caption. Figure taken from [Xu+15].

As datasets grew and computational capabilities expanded, image captioning models began to incorporate more structured and semantically informed mechanisms. The Bottom-Up and Top-Down Attention model [And+18] marked a key milestone by combining two complementary stages: a bottom-up mechanism that uses an object detector such as Faster R-CNN [Ren+15] to extract region-level visual features, and a top-down decoder that adaptively attends over these regions during caption generation. This dual-level architecture enabled more accurate object grounding and improved compositional reasoning, and it quickly became a standard in captioning benchmarks such as MSCOCO [Lin+14].

With the rise of transformers [Vas+17], modern image captioning systems have increasingly adopted fully transformer-based architectures. These models, including OSCAR [Li+20], VinVL [Zha+21], and BLIP [Li+22], leverage large-scale pre-training on image-text pairs and integrate both vision and language processing in unified frameworks. They commonly use visual encoders, such as ViT or ResNet, to extract image features, which are then fused with text representations via multimodal transformers. This integration supports both contrastive alignment and conditional generation, enabling strong generalization and zero-shot transfer across tasks.

A representative model of this paradigm is Bootstrapping Language-Image Pre-training (BLIP) [Li+22], which introduces a unified vision-language framework combining three objectives: Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM). As shown in Figure 2.9, BLIP consists of an image encoder, a text encoder, and a transformer-based text decoder with cross-attention over image features. The ITC module (left) aligns image and text embeddings by projecting them into a shared feature space and applying a contrastive loss that pulls paired embeddings together while pushing apart mismatched ones, enabling retrieval. The ITM module (middle) enforces

fine-grained alignment by treating image–text correspondence as a binary classification problem: the image-grounded text encoder, which incorporates cross-attention over visual features, predicts whether a caption matches its image. The LM module (right) supports caption generation, where the image-grounded text decoder uses causal self-attention to predict each token based on preceding tokens and the visual context, producing coherent descriptions. Together, these modules allow BLIP to combine discriminative learning for retrieval and alignment (via ITC and ITM) with generative modeling for captioning (via LM), yielding a flexible architecture that supports both understanding and generation tasks.

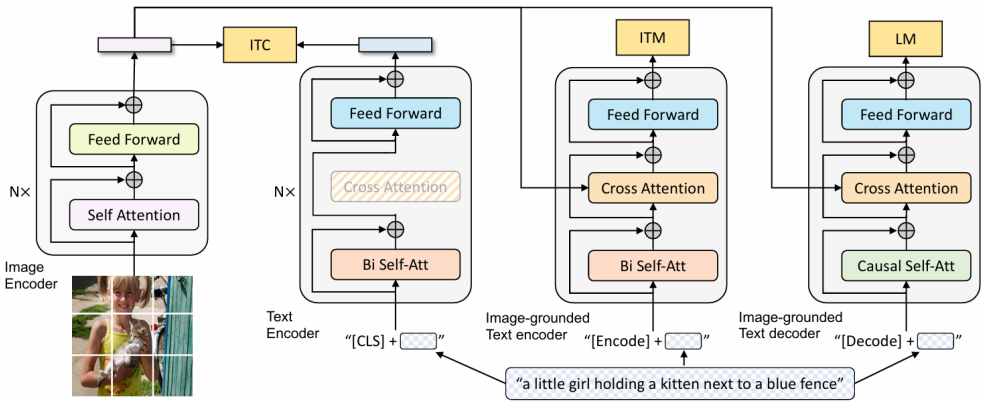


Fig. 2.9: Overview of the BLIP architecture [Li+22], which integrates Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM) through a unified vision-language transformer framework. Each module dynamically combines image and text features via cross-attention to support both retrieval and generation tasks. Figure taken from [Li+22]

To further improve scalability and reasoning capabilities, BLIP was extended to BLIP-2 [Li+23], which decouples visual and language components by using a frozen image encoder and a large pre-trained language model, connected through a lightweight Querying Transformer (Q-Former), as later depicted in Figure 3.1. This modular design facilitates compatibility with powerful LLMs such as Flan-T5 or Vicuna, while significantly reducing training costs. Building on this foundation, InstructBLIP [Dai+23] introduces instruction tuning to better align image captioning with user intent and application goals. The details of InstructBLIP and its usage in this thesis are presented in Chapter 3.

Instruction-tuned systems like InstructBLIP act as precursors to a broader class of Multimodal Large Language Models (MLLMs). These models unify visual understanding and language generation by pairing frozen visual encoders with powerful LLMs, connected through an intermediate projection or adapter module. A general architecture of such MLLMs is illustrated in Figure 2.10. The image is first processed by a visual encoder that extracts high-dimensional visual features. These features are then passed through an adapter module, which projects them into the language embedding space. The adapted features are concatenated or integrated with a textual prompt and forwarded to the LLM, which produces a grounded and context-aware response. This framework supports a

wide range of multimodal tasks beyond captioning, such as visual question answering, where the model generates answers to natural-language queries about an image; dialogue grounding, where conversational utterances are linked to the relevant visual content; and object referencing, where the model identifies or locates a specific object in an image based on a textual description. Crucially, the adapter-based approach enables scalable training and broad reuse of powerful LLMs without requiring full joint tuning, making it a key architectural design in state-of-the-art MLLMs.

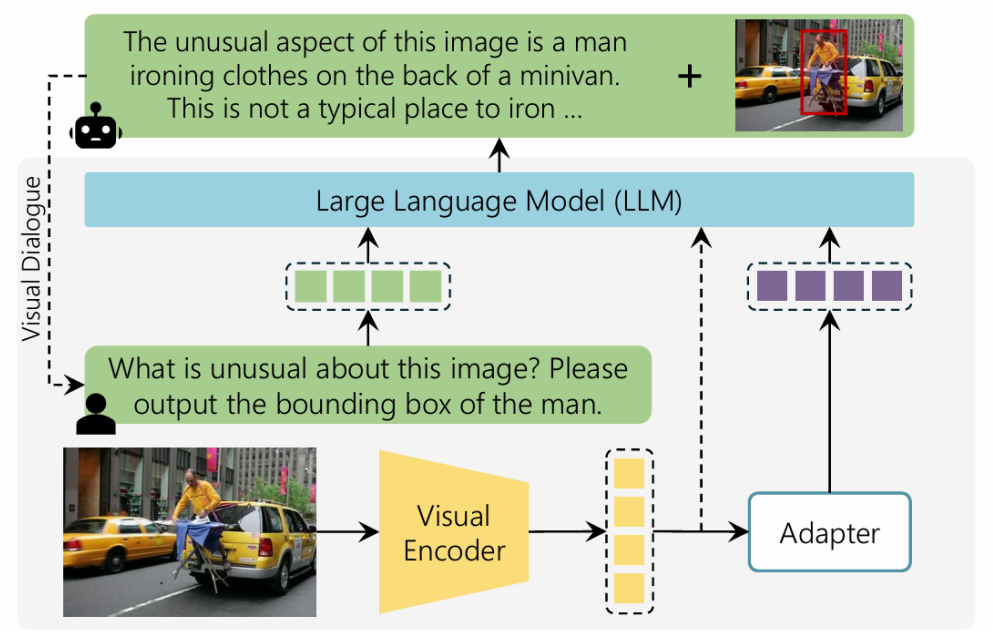


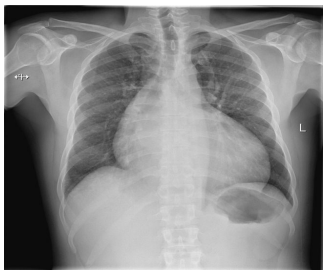
Fig. 2.10: General architecture of Multimodal Large Language Models (MLLMs). The image is encoded via a vision encoder, passed through an adapter to align modalities, and then injected into a frozen language model to generate grounded textual responses. Figure adapted from [Caf+24].

2.4 Diagnostic Captioning

Diagnostic Captioning (DC) is the task of automatically generating clinically accurate and contextually appropriate textual descriptions from medical images, most commonly from modalities such as chest X-rays, CT, or MRI. In contrast to generic image captioning, which aims to describe visible content in everyday language, diagnostic captioning requires models to identify medically relevant abnormalities, use phrasing consistent with radiological reporting conventions, and avoid hallucinations or unsubstantiated conclusions. The generated text must therefore capture not only the visual findings but also their clinical significance, including correct use of negation, expressions of uncertainty, and domain-specific terminology.

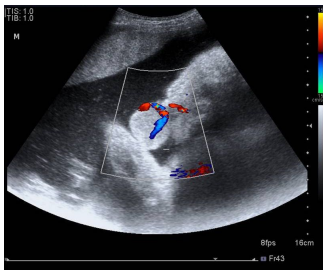
As a research task, diagnostic captioning resides at the intersection of computer vision, medical image analysis, and clinical natural language generation. It plays a pivotal role in

applications such as automated radiology reporting, clinical decision support systems, and standardized medical documentation, particularly in scenarios where diagnostic throughput, consistency, or expert availability are constrained. Importantly, these systems are not intended to replace human experts, but rather to augment their workflow by producing preliminary drafts, highlighting salient visual findings, and promoting standardization across large volumes of imaging studies. Figure 2.11 presents three representative examples drawn from the ImageCLEFmedical dataset [Rüc+24a]. Each image is accompanied by its ground truth caption, reflecting the kind of clinically focused, modality-specific descriptions expected in diagnostic captioning.



(a) "Chest X-ray showing enlarged cardiac silhouette with cardiothoracic ratio of 70%, and mild pulmonary congestion."

(b) "Contrast-enhanced CT image shows 3.6 × 4.5 × 3.1 cm well defined heterogeneous enhancing mass in paraaortic space (arrow)."



(c) "Ultrasound evaluation with color Doppler, showing a mass protruding from the mouth with a branching pattern of the feeder vessels."

Fig. 2.11: Representative examples of diagnostic image-caption pairs from the ImageCLEFmedical dataset [Rüc+24a]. The captions shown are ground truth annotations provided in the dataset and illustrate clinically relevant findings across multiple imaging modalities.

Early models for diagnostic captioning followed the encoder–decoder paradigm established in generic captioning. These systems typically combined CNNs such as ResNet [He+16] for visual encoding with RNNs or LSTMs [HS97] for text generation. Attention mechanisms, introduced in follow-up work, allowed models to dynamically focus on relevant image regions while generating tokens, improving interpretability and localization. However, these models often struggled with domain-specific challenges such as subtle findings, ambiguous phrasing, and clinical correctness, limiting their practical deployment.

A second wave of research in diagnostic captioning has increasingly leveraged structured modeling and auxiliary supervision to enhance the factual correctness and conceptual consistency of generated radiology reports. Here, structured modeling refers to methods that explicitly represent relationships between clinical concepts, such as diseases, anatomical locations, and observations, rather than treating report generation as a purely sequence-to-sequence task. Similarly, semantic alignment in this context refers to ensuring that the textual content of the generated report correctly corresponds to the visual information extracted from the medical images, aligning image features with the appropriate clinical concepts. For instance, Jing et al. [JXX18] introduced a hierarchical generation model guided by predicted clinical tags, showing that auxiliary tasks, such as tag prediction, can anchor report content in medically relevant concepts and improve the consistency

of the generated text. Other work incorporates explicit constraints into the generation process to reduce factual errors. Specifically, attention modulation refers to adjusting the model's focus on different regions of the input image based on clinical context, while knowledge-constrained generation means restricting the output to conform to known medical relationships or rules, preventing implausible combinations of findings. For example, Han et al. [Han+20] propose a neural-symbolic framework that integrates domain knowledge graphs, structured representations of medical entities (e.g., diseases, symptoms, anatomy) and their relationships, into the generation process. These graphs enforce logical consistency and improve factual accuracy in the reports. More advanced approaches improve robustness by explicitly aligning image-derived visual features (features extracted from X-ray images using a convolutional network) with diagnostic labels (e.g., "pneumonia" or "cardiomegaly") to ensure that the learned visual representations correspond to clinically meaningful concepts. Yang et al. [Yan+23] further enhance this approach by using a learned knowledge base, which encodes associations between diseases, imaging findings, and textual descriptions in a structured embedding space. This knowledge base is constructed from training data and captures statistical co-occurrences and relationships among clinical entities, guiding the model to generate reports that are both semantically consistent and clinically accurate. Their framework also leverages multiple associated reports per image to improve generalization across diverse report styles. Collectively, these methods demonstrate the effectiveness of domain-specific supervision, particularly when applied to radiology-focused datasets such as MIMIC-CXR [Joh+19] and OpenI [Dem+16], which have become standard benchmarks in the field.

Recent advances in multimodal learning have led to the emergence of more powerful and flexible diagnostic captioning systems. Transformer-based VLMs such as GIT-CXR [Sîr+25], as well as Flamingo-based architectures like Med-Flamingo [Moo+23], have achieved strong performance by combining pretraining on large-scale image–text pairs with fine-tuning on radiological data. More recently, instruction-tuned architectures such as Med-Gemma, an open vision-language model suite trained on paired medical images and text across multiple modalities, enable better alignment with clinical tasks including captioning, classification, and visual question answering.[2] Leading Multimodal LLMs (MLLMs) like Med-Gemini [Saa+24], which integrates a large language model with medical image understanding through visual adapters and multimodal instruction tuning, support few-shot and zero-shot generalization across medical imaging and reporting tasks [YX+24]. These models represent a shift toward flexible, clinically aligned captioning systems that balance linguistic fluency, factual accuracy, and task-specific adaptability.

[2] https://medgemma.org/

3 Implemented Methods and Systems

This chapter presents the methods and systems developed in the context of this thesis. Each approach is categorized under one of the three main post-training paradigms, Supervised Fine-Tuning, Reinforcement Learning, and Test-Time Scaling, based on its primary mechanism, although some methods may incorporate elements from more than one category. The systems detailed in Sections 3.1, 3.3, and 3.4 form the core of our participation in the Caption Prediction sub-task of the ImageCLEFmedical 2025 challenge [Cha+25; Dam+25]. These submissions led to a 5th place ranking among 8 competing research teams. Additional systems, introduced in the remaining sections, were explored in the later stages of this thesis, after the ImageCLEFmedical challenge, to further evaluate complementary strategies and design variations.

3.1 Instruction Fine-Tuning with InstructBLIP

Instruction fine-tuning, as explained in Section 2.3, refers to the process of training a language model on curated pairs of instructions (prompts) and expected outputs (responses). Unlike standard supervised learning that may rely on task-specific training examples, instruction fine-tuning teaches the model to generalize across a wide range of tasks by following natural language instructions. This technique, originally developed in the NLP community [Chu+24], has been shown to improve zero-shot and few-shot generalization across diverse task types such as summarization, question answering, and classification. Instruction-tuned models like FLAN [Wei+22a] and Alpaca[1] have demonstrated strong capabilities to align with user intent by virtue of their enhanced instruction-following behavior.

InstructBLIP [Dai+23] builds upon the BLIP-2 [Li+23] architecture, introduced in Section 2.3, by extending it to the instruction-tuning setting. As shown in Figure 3.1, it integrates three key components: a frozen image encoder (ViT-g/14), a frozen large language model (either FlanT5 or Vicuna), and a trainable Query Transformer (Q-Former) that bridges the vision and language modalities. A central feature of the Q-Former is its use of learnable queries: trainable vectors that attend to the image encoder's output via cross-attention. These queries act as an adaptive interface that extracts task-relevant visual

[1] https://crfm.stanford.edu/2023/03/13/alpaca.html

features, in contrast to static pooling operations. In InstructBLIP, the input instruction is injected into the Q-Former's attention layers so that the queries adapt their focus depending on the task. This design makes the visual features instruction-aware, ensuring that the downstream language model receives representations aligned with the user's intent. The resulting features are linearly projected and used as soft prompts to condition the frozen LLM. Training proceeds by updating only the Q-Former (including the queries) while keeping both the image encoder and the language model frozen, using a large collection of instruction–response pairs across multiple vision-language tasks.

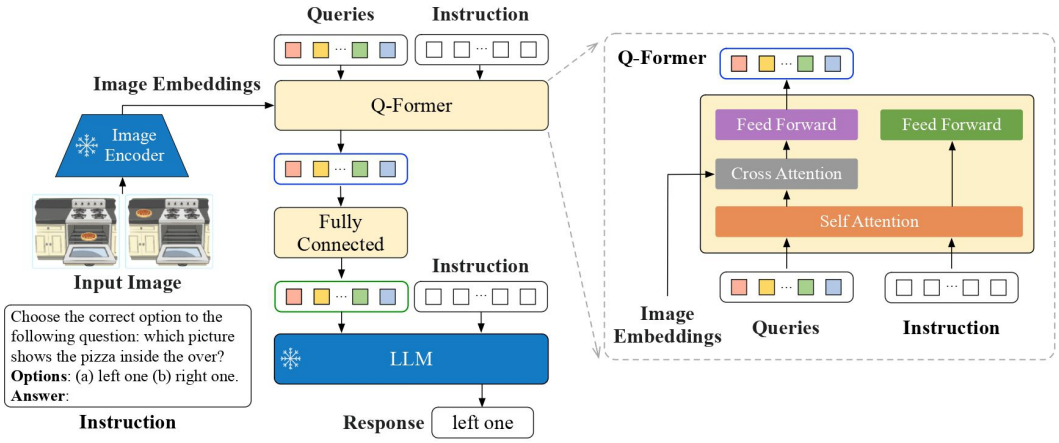


Fig. 3.1: Architecture of InstructBLIP [Dai+23]. Visual features are extracted from a frozen image encoder via a Query Transformer (Q-Former) that receives both learnable queries and the task instruction. These instruction-aware features are projected and injected into a frozen LLM, which generates the response. Figure adapted from [Dai+23].

In this thesis, InstructBLIP is adapted to the medical domain by fine-tuning it on an instruction-formatted variant of the ImageCLEFmedical 2025 captioning dataset [Rüc+24a]. Consistent with the original setup, both the vision encoder (ViT-g/14) and the large language model (FlanT5) remain frozen, while only the Q-Former, and its learnable queries, is updated during training. Each training sample is paired with a handcrafted instruction prompting the model to produce a concise and clinically accurate description of the input radiology image. This setup enables the model to leverage the instruction-following capabilities of the LLM while grounding its output in visual features that are adapted to the clinical task. The resulting system serves as our primary and top-performing submission to the ImageCLEFmedical 2025 Caption Prediction sub-task and is further evaluated in Chapter 5.

3.2 Contrastive Fine-Tuning with InfoNCE

While standard instruction fine-tuning optimizes a model to generate accurate captions using cross-entropy loss, it does not explicitly enforce the ability to distinguish correct captions from incorrect ones. As a result, VLMs may still produce plausible-sounding yet

clinically inaccurate outputs, particularly in cases involving subtle visual cues or prior biases in the training data.

To address this limitation, InstructBLIP is extended in this work with a contrastive learning objective that promotes fine-grained alignment between radiology images and their corresponding reports. This extension introduces an additional InfoNCE-style loss [OLV18] during fine-tuning, encouraging matched image–text pairs to be close in a shared embedding space while simultaneously repelling mismatched pairs.

In practice, image embeddings are extracted from the frozen vision encoder and text embeddings from the frozen language model's encoder (FlanT5). These embeddings are projected into a shared 512-dimensional space using separate trainable linear projection layers for text and images. The contrastive loss is then computed using a symmetric Information Noise-Contrastive Estimation (InfoNCE) formulation over all pairs within each training batch:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, t_j)/\tau)} - \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_j, t_i)/\tau)} \right], \tag{3.1}$$

where v_i and t_i are the projected visual and textual embeddings for the i-th sample in the batch, $\text{sim}(\cdot)$ denotes cosine similarity, τ is a temperature parameter, and N is the batch size.

The overall training objective combines this contrastive loss with the standard captioning loss $\mathcal{L}_{\text{captioning}}$ (cross-entropy over generated tokens):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{captioning}} + \lambda \cdot \mathcal{L}_{\text{contrastive}}, \tag{3.2}$$

where λ is a tunable hyperparameter controlling the influence of the contrastive term. A value of $\lambda = 0.2$ was selected empirically based on performance on a held-out development set.

This joint optimization requires no additional supervision or curated negatives. By leveraging in-batch negatives, i.e., unmatched captions and images from other samples in the batch, the model learns to associate clinically appropriate descriptions with the correct visual inputs and to reject inappropriate or misleading ones. This setup encourages greater factual grounding, reduces hallucinated findings, and improves discrimination between visually similar but clinically distinct cases.

The contrastive module is implemented with minimal architectural modification and is trained concurrently with the instruction-finetuned InstructBLIP. Experimental results

presented in Chapter 5 demonstrates that this training strategy improves performance on alignment-related metrics, such as Image-Caption Similarity (Section 5.1.4) and AlignScore (Section 5.1.5), in the generated reports for the ImageCLEFmedical 2025 captioning task.

3.3 Reinforcement Signal–Driven Training with Mixer

Mixer is a training strategy that augments traditional cross-entropy optimization with a task-specific reinforcement signal. The goal is to directly improve evaluation-time metrics, such as BERTScore for semantic similarity or UMLS Concept F1 for clinical factuality, by rewarding model outputs that better align with downstream objectives. These metrics, described in detail in Section 5.1, are often non-differentiable and cannot be optimized through standard supervised losses alone. While the overall training pipeline remains supervised in structure, the inclusion of a learned reward signal based on model predictions places this method within the broader family of reinforcement learning approaches.

The Mixer approach builds on Self-Critical Sequence Training (SCST) [Ren+17b], a reinforcement learning technique designed specifically for sequence generation tasks. Unlike traditional RL setups that require external critics or value networks, SCST leverages the model's own predictions as a dynamic baseline for computing rewards. This self-referential setup reduces variance and simplifies optimization. In practice, the model generates two captions per training instance: a greedy caption $\hat{y}$ via deterministic decoding, and a sampled caption y^s using stochastic decoding strategies such as top-p sampling or diverse beam search. These outputs are then compared using a reward function tailored to the specific goals of diagnostic captioning, computed as a combination of six task-relevant metrics that each produce scores in the range [0, 1]. Four of these, BERTScore (Section 5.1.1), ROUGE-1 (Section 5.1.2), BLEURT (Section 5.1.3), and Image–Text Similarity (Section 5.1.4), focus on semantic relevance, while the remaining two, AlignScore (Section 5.1.5) and UMLS Concept F1 (Section 5.1.6), emphasize clinical factuality. To obtain the final reward, the scores from the relevance-based metrics are first averaged, as are those from the factuality-based metrics. The overall reward is then computed as the mean of these two intermediate averages.

The reinforcement signal is defined in terms of the *advantage*, which quantifies how much better (or worse) the sampled caption y^s performs relative to the greedy baseline $\hat{y}$ according to the composite reward function:

$$Adv(y^s) = r(y^s) - r(\hat{y}), \tag{3.3}$$

where $r(y)$ denotes the scalar reward computed using the six evaluation metrics, following the aggregation procedure outlined above. A positive advantage indicates that the sampled caption outperforms the greedy one in terms of overall alignment with the task-specific evaluation criteria, while a negative value implies the opposite.

As mentioned earlier, the training objective is not limited to the standard cross-entropy loss, which in its simplest form is defined as $-\sum_i y_i \log p_i$, where y_i is the true label (one-hot) and p_i is the predicted probability. In the sequence setting considered here, this becomes:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, \mathcal{I}), \tag{3.4}$$

where $\pi_\theta(y_t \mid y_{<t}, \mathcal{I})$ denotes the probability of the t-th token given the previous tokens and the input image $\mathcal{I}$.

This is combined with a reinforcement learning loss defined using the SCST formulation:

$$\mathcal{L}_{\text{RL}} = -Adv(y^s) \cdot \log \pi_\theta(y^s), \tag{3.5}$$

where $\pi_\theta(y^s)$ is the likelihood of the sampled caption y^s, and $Adv(y^s)$ is the advantage score computed as in Equation 3.3.

The final training objective combines both terms into a weighted sum:

$$\mathcal{L}_{\text{total}} = (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{RL}}, \tag{3.6}$$

where α is a mixing coefficient that controls the balance between supervised and reinforcement learning signals.

To ensure stable optimization, the contribution of the reinforcement signal is introduced progressively throughout training. This is achieved by linearly increasing the mixing coefficient α over time, according to the following schedule:

$$\alpha(e) = \alpha_{\text{max}} \cdot \frac{e}{E}, \tag{3.7}$$

where e denotes the current epoch, E is the total number of training epochs, and α_{max} is the maximum reinforcement weight (typically set close to 1.0). This gradual ramp-up allows the model to first focus on learning fluent and well-formed captions via cross-entropy, and then progressively shift toward optimizing task-specific evaluation metrics through reinforcement learning.

3.4 Test-Time Caption Reranking with MedCLIP

To improve the quality of generated image captions at inference time, a test-time reranking strategy is proposed that operates independently of the underlying captioning model. This method can be applied on top of any pretrained captioning backbone, such as InstructBLIP (Section 3.1), and serves as a post-processing step that selects the most visually grounded caption among a set of candidates. The reranking mechanism relies on MedCLIP [Wan+22], a contrastive vision-language model trained specifically on radiological data.

The architecture and workflow of MedCLIP are illustrated in Figure 3.2, which highlights how clinical knowledge is used to supervise the alignment of image and text representations. This method builds upon the foundation of CLIP [Rad+21], first introduced in Section 2.1, which jointly trains a visual encoder and a text encoder to align paired image and text embeddings in a shared latent space. CLIP is trained on 400 million natural image-text pairs from the internet using a contrastive loss that pulls matched pairs closer while pushing mismatched ones apart. Although CLIP has demonstrated remarkable zero-shot capabilities across general vision-language tasks, its direct application to the medical domain is hindered by the scarcity of large-scale paired image-report datasets and the subtle, fine-grained semantics of clinical language.

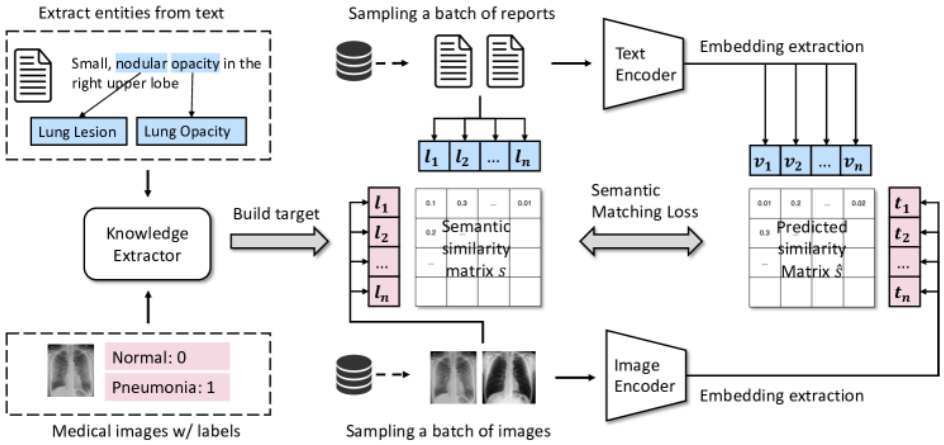


Fig. 3.2: Overview of the MedCLIP architecture. Medical entities are extracted from unpaired images and reports to build a semantic similarity matrix. The model uses this signal to contrastively train aligned visual and textual embeddings, which are later used to rerank caption candidates based on image-text similarity. Figure taken from [Wan+22]

To address these challenges, MedCLIP adapts the CLIP framework by introducing two key modifications. First, it decouples the contrastive training process to allow learning from unpaired data. Radiological datasets often contain image-only or text-only samples, and strict pairing is expensive and limited. To overcome this, MedCLIP constructs a semantic similarity matrix between sampled images and texts by extracting medical concepts from both modalities and aligning them to a common ontology. On the text side, clinical entities are extracted from reports using tools such as MetaMap, which links terms to the

Unified Medical Language System (UMLS). On the image side, diagnostic labels provided in radiology datasets (e.g., "pneumonia", "cardiomegaly") are likewise mapped to UMLS concepts through their standardized codes. This entity-level alignment ensures that both modalities are represented in the same semantic space. The aligned entities are then converted into multi-hot vectors, and their cosine similarity provides the soft supervision signal for contrastive training.

Second, MedCLIP replaces the standard InfoNCE loss with a semantic matching loss that reduces false negative noise. In medical datasets, different patient cases may express the same findings; treating them as negatives can harm representation learning. Instead, the semantic similarity between any image-text pair is used to compute soft targets, and contrastive learning is guided by cross-entropy between these targets and predicted cosine similarities. Formally, let v_i and t_j denote the normalized visual and textual embeddings of the i-th image and j-th text in a batch, respectively. The semantic similarity between their corresponding medical entity labels l_{img} and l_{txt} defines a soft target distribution y_{ij}. The predicted similarity is obtained from the cosine score $s_{ij} = v_i^\top t_j$. The semantic matching loss is then given by:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} \log \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{N} \exp(s_{ik}/\tau)}, \tag{3.8}$$

where τ is a temperature hyperparameter and N is the batch size. The final training objective symmetrizes this loss over both image-to-text and text-to-image directions [Wan+22]. The model consists of a vision encoder (e.g., ResNet-50 [He+16] or Swin Transformer [Liu+21]) and a text encoder (e.g., BioClinicalBERT[2], each followed by a projection head to produce embeddings in a shared space. These embeddings are normalized and used for computing similarities during both training and inference.

At test time, the reranking module operates as follows: for a given image, the captioning model generates a set of $m = 4$ candidate captions using beam search. Each caption is encoded by MedCLIP and scored based on its similarity to the image embedding. The caption with the highest similarity is selected as the final output. This strategy enhances the clinical plausibility of the generated description by prioritizing captions that are not only syntactically fluent but also semantically aligned with the visual content.

[2] `https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT`

4 Data

This chapter provides general background information along with an exploratory data analysis of the ImageCLEFmedical 2025 [Dam+25] dataset. This dataset served as the primary benchmark for evaluating the performance and effectiveness of the systems developed in this thesis, which were presented in Chapter 3.

4.1 ImageCLEFmedical 2025

The ImageCLEFmedical 2025 [Dam+25] dataset is based on an extended version of the Radiology Objects in COntext Version 2 (ROCOv2) [Rüc+24a] and serves as the foundation for all three sub-tasks of the challenge, described in Chapter 1. It comprises 97,368 medical images from various imaging modalities, each accompanied by a corresponding diagnostic caption and a set of medical concepts represented as Unified Medical Language System (UMLS) [Bod04] terms. As this thesis focuses on the Caption Prediction sub-task, the analysis presented here will be limited to that part of the dataset.

The dataset was initially provided in two official splits: a training set and a validation set, containing 80,091 and 17,277 images, respectively. However, all methods discussed in Chapter 3 were evaluated using a custom split, created by merging the original sets and repartitioning them into three subsets: training, validation, and development (or private test) sets, using a 75%-10%-15% ratio. This new partitioning was performed using stratification based on both concept distribution and caption length. The effectiveness of the stratification was confirmed by visualizing and comparing the distributions across the new splits.

4.2 Caption Prediction

The Caption Prediction sub-task focuses on generating coherent and clinically relevant textual descriptions for medical images. Each image in the dataset is paired with a single caption. In total, the dataset contains 97,268 captions, out of which 96,866 are unique, resulting in a uniqueness rate of 99,48%. Caption lengths vary considerably: the shortest caption consists of a single word, while the longest reaches up to 778 words. On average, a caption contains approximately 21 words. This variation suggests that while many

captions describe routine imaging procedures, others provide more detailed and specific clinical observations. Table 4.1 summarizes the key statistics related to caption frequency and length.

Statistic	Value
Total Captions	97,268
Unique Captions	96,866
Percentage Unique	99.48%
Minimum Length	1 word
Maximum Length	778 words
Average Length	21 words

Tab. 4.1: Descriptive statistics of captions in the dataset.

The distribution of caption lengths is highly skewed, with most captions being relatively short and only a few extending to several hundred words. This pattern is illustrated in Figure 4.1, which presents two histogram plots. The left subplot shows the full distribution of caption lengths across the dataset, while the right subplot provides a zoomed-in view limited to captions containing fewer than 200 words. This visualization highlights the presence of a long-tail distribution, with the vast majority of captions concentrated in the lower range of lengths.

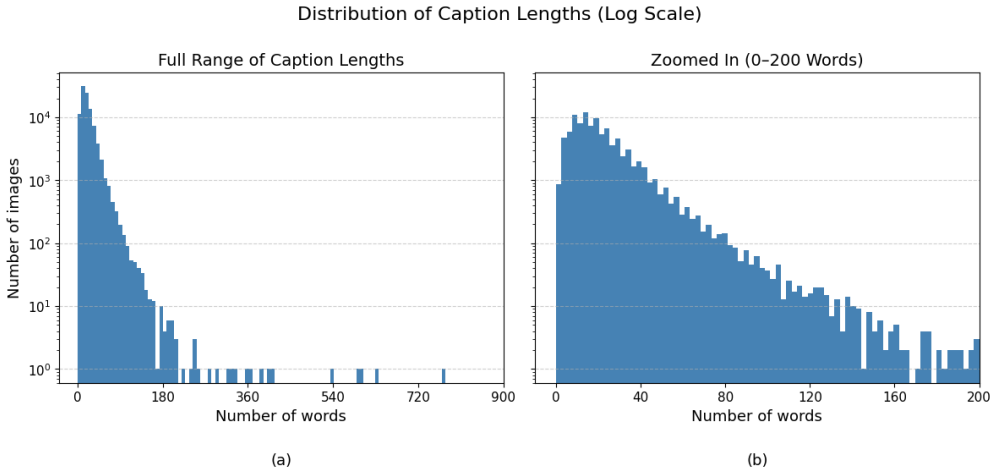


Fig. 4.1: Distribution of caption lengths (in number of words) on a logarithmic scale. The left plot (a) shows the full range of caption lengths across the dataset, while the right plot (b) provides a zoomed-in view limited to captions with up to 200 words.

To better understand common phrase patterns in the dataset, a frequency analysis of bigrams and trigrams was performed. These n-grams represent pairs and triplets of consecutive words that frequently appear across captions. Table 4.2 presents the ten most common bigrams and trigrams. The results show a strong presence of modality-related and anatomical expressions, reflecting the specialized vocabulary typical in radiology reporting. Additionally, several frequent phrases involve visual markers (e.g., “white arrow”, “red

arrow") or generic reporting verbs (e.g., "showing"), which are commonly used to draw attention to specific findings or regions of interest within the image.

Bigram	Freq.	Trigram	Freq.
computed tomography	10,539	magnetic resonance imaging	2,777
ct scan	7,381	computed tomography scan	2,042
chest xray	4,148	computed tomography ct	1,444
magnetic resonance	4,027	ct scan showing	1,357
white arrow	3,328	chest xray showing	1,041
red arrow	2,842	ct scan abdomen	767
resonance imaging	2,836	ct computed tomography	749
tomography scan	2,292	chest computed tomography	691
scan showing	2,275	abdominal computed tomography	498
image showing	2,174	computed tomography image	633

Tab. 4.2: Top 10 most frequent bigrams and trigrams in the captions.

To complement the n-gram analysis, a word cloud visualization was generated to highlight the most frequently occurring non-stopwords in the dataset. As shown in Figure 4.2, prominent terms such as "computed tomography", "ct scan", "white arrow", and "showing" dominate the visualization. These frequent expressions reflect both the procedural nature of medical imaging and the emphasis on anatomical references and visual indicators within the captions. The word cloud provides an intuitive overview of the lexical patterns in the dataset, visually emphasizing domain-specific terminology.

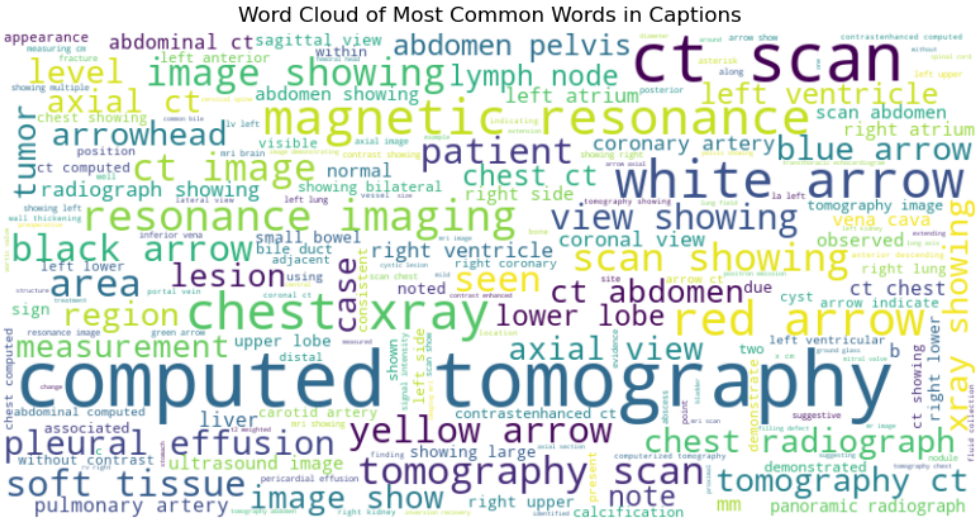


Fig. 4.2: Word cloud of the most frequent non-stopwords appearing in the captions. Word size is proportional to frequency.

Finally, it is important to note that all captions are subjected to a standardized preprocessing pipeline prior to evaluation, as specified by the task organizers. This ensures consistent and fair comparison across systems. The preprocessing includes:

- Converting all characters to lowercase,
- Replacing numeric values with their word equivalents (e.g., “10” becomes “ten”), and
- Removing punctuation marks.

These steps reduce superficial variation in model outputs and shift evaluation focus toward semantic accuracy and fluency.

5 Experiments and Results

This chapter outlines the experimental framework followed in this thesis, including the evaluation methodology and performance analysis of the diagnostic captioning systems. It also details the participation of the AUEB NLP Group in the ImageCLEFmedical 2025 [Dam+25] competition and the metrics used to assess model effectiveness.

5.1 Evaluation Metrics

The evaluation metrics used in this thesis align with those employed in the ImageCLEFmedical 2025 competition [Dam+25]. These metrics are designed to assess the quality of generated captions from multiple perspectives. Specifically, they are grouped into two main categories: relevance, which evaluates how well the caption matches the image and the reference text, and factuality, which focuses on the clinical accuracy and consistency of the generated content. Together, these complementary dimensions offer a comprehensive framework for evaluating the effectiveness of diagnostic captioning systems.

5.1.1 BERTScore

BERTScore [Zha+20] is an automatic evaluation metric for text generation that measures the similarity between a generated (candidate) sentence and a reference sentence using contextualized embeddings from pretrained language models. Unlike traditional lexical overlap-based metrics such as ROUGE [Lin04] (discussed in Section 5.1.2), BERTScore captures semantic similarity by leveraging token-level embeddings from Transformer models like BERT [Dev+19].

To compute the BERTScore between a reference sentence x and a candidate sentence $\hat{x}$, each token in both sequences is embedded using a contextualized language model—in this case, the microsoft/deberta-xlarge-mnli[1] model. The token embeddings are then compared pairwise using cosine similarity. For each token in $\hat{x}$ (the candidate), the most similar token in x (the reference) is identified, and vice versa. These maximum similarities are aggregated to produce the final score. Figure 5.1 illustrates this process, showing how

[1] https://huggingface.co/microsoft/deberta-v2-xlarge-mnli

contextual embeddings are used to construct a similarity matrix, from which the strongest alignment paths are extracted.

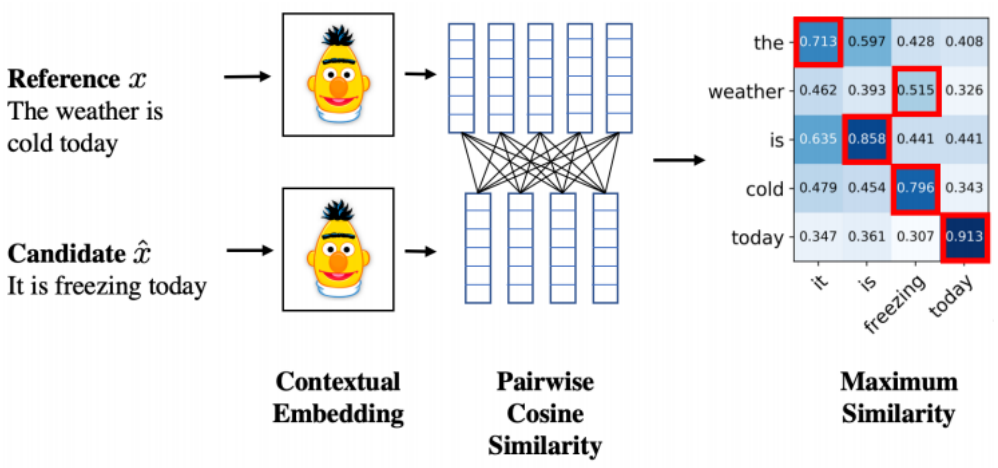


Fig. 5.1: Illustration of BERTScore computation between a reference sentence x (top) and a candidate sentence $\hat{x}$ (bottom). Tokens are embedded using a Transformer model, and pairwise cosine similarity scores are computed. Maximum similarity scores are selected per token for aggregation. Figure taken from [Zha+20].

The original BERTScore framework [Zha+20] proposes three scoring variants: Precision (P), Recall (R), and F1 (F), depending on the directionality of token alignment between the candidate and reference. In the context of this thesis, we adopt the Recall variant with Inverse Document Frequency (IDF) weighting, which prioritizes semantic coverage of the reference content by the candidate. IDF is a common term-weighting scheme in information retrieval that assigns higher weights to rare, content-rich tokens and lower weights to frequent ones. This choice aligns with the official evaluation setup of the ImageCLEFmedical 2025 competition [Dam+25].

The $\mathrm{R}_{\text{BERT-IDF}}$ (Recall BERTScore with IDF weighting) is computed as follows:

$$\mathrm{R}_{\text{BERT-IDF}} = \frac{\sum_{i=1}^{m} \mathrm{IDF}(x_i) \cdot \max_{1 \le j \le n} \mathrm{cosine}\big(E(x_i), E(\hat{x}_j)\big)}{\sum_{i=1}^{m} \mathrm{IDF}(x_i)}. \tag{5.1}$$

Here, $\mathbf{x} = (x_1, \ldots, x_m)$ and $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$ denote the sequences of tokens in the reference and candidate sentences, respectively. $E(x_i)$ and $E(\hat{x}_j)$ represent the contextual embeddings of the corresponding tokens, and the cosine similarity measures their semantic closeness. The $\max$ operator selects, for each reference token, the candidate token with the highest similarity, ensuring that the most relevant matches contribute to the score. The IDF weights are computed from the test corpus and serve to emphasize rare and content-rich tokens while down-weighting common ones. To obtain the overall BERTScore for a model, the recall score is first calculated independently for each caption-reference pair in the

evaluation set. The final metric is then derived by averaging these individual scores across all examples in the dataset.

While BERTScore has the advantage of capturing semantic similarity, it also has important limitations, particularly in domains like medicine where factual correctness is critical. One major flaw is that BERTScore may assign a high similarity score to fluent but factually incorrect outputs. For example, consider the following two sentences:

- **Reference:** "There is no evidence of pneumonia in the right lower lobe."
- **Candidate:** "There is evidence of pneumonia in the right lower lobe."

Despite the crucial difference introduced by the word "no", BERTScore might still assign a high similarity score due to the strong semantic overlap in surrounding tokens. This insensitivity to negation or factual contradiction is a proven weakness [HB21] and makes BERTScore unsuitable as a standalone metric in clinical applications. Additionally, because BERTScore relies on large pretrained models, it can be computationally expensive, especially when evaluating large test sets.

Overall, while BERTScore contributes valuable insight into semantic fidelity, it must be complemented by factuality-oriented metrics to ensure reliability in medical captioning tasks.

5.1.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin04] is a set of metrics commonly used to evaluate automatic text summarization and generation systems by measuring the lexical overlap between generated outputs and reference texts. It is particularly useful when exact wording matters, and it has been widely adopted due to its simplicity and interpretability.

Several variants of ROUGE exist, including ROUGE-N (n-gram overlap), ROUGE-L (longest common subsequence), and ROUGE-W (weighted variant of the latter). In this thesis, and in the ImageCLEFmedical 2025 competition [Dam+25], we focus on ROUGE-1 (F-measure), which evaluates the overlap of unigrams (i.e., individual words) between the generated caption and its reference.

Formally, ROUGE-1 F-measure is defined as:

$$\text{ROUGE-1}_F = \frac{2 \cdot P \cdot R}{P + R} \tag{5.2}$$

where precision P and recall R are given by:

$$P = \frac{\text{\# overlapping unigrams}}{\text{\# unigrams in candidate}}, \quad R = \frac{\text{\# overlapping unigrams}}{\text{\# unigrams in reference}} \tag{5.3}$$

The final ROUGE-1 score is computed by averaging the F-measure across all caption-reference pairs in the evaluation set.

While ROUGE-1 is effective in capturing surface-level similarity and penalizing missing or extraneous words, it does not account for semantic similarity or paraphrasing. For instance, it would assign a low score to the pair "chest radiograph" and "x-ray of the chest", despite them being semantically equivalent.

5.1.3 BLEURT

BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) is a learned metric for evaluating the quality of text generation systems [SDP20]. Unlike traditional n-gram-based metrics such as BLEU [Pap+02] and ROUGE [Lin04], BLEURT uses pretrained language models and fine-tuning on human-annotated quality scores to capture semantic similarity and linguistic fluency in a way that better aligns with human judgment.

At its core, BLEURT is a regression model built on top of BERT [Dev+19], which is fine-tuned using reference-candidate sentence pairs and their associated human ratings. The model is trained to predict these scores directly, enabling it to approximate subjective quality assessments. BLEURT computes a scalar score for a candidate-reference pair, indicating the predicted human rating. The predicted scores are typically in the range $[-1, 1]$, with higher values corresponding to better quality. This range reflects the normalization of human ratings used during BLEURT's fine-tuning on quality estimation datasets.

A distinguishing feature of BLEURT is its two-phase training procedure. First, it undergoes pretraining on large-scale synthetic sentence pairs generated by perturbing Wikipedia sentences using techniques such as masked language modeling, back-translation (translating a sentence to another language and back to create a paraphrase [SHB16]), and random word dropping. During this pretraining, the model learns to predict a continuous similarity score for each sentence pair, which reflects the degree of semantic change introduced by the perturbation, allowing it to capture fine-grained differences between sentences.

This step improves generalization to unseen domains. Next, BLEURT is fine-tuned on real human-labeled data from quality estimation benchmarks such as the WMT Metrics Shared Tasks [Boj+17].

In the context of this thesis, BLEURT is employed using the official BLEURT-20 checkpoint, which was pretrained on synthetic data and fine-tuned on human ratings. The BLEURT score is calculated for each caption, and the final score is the mean over all image-caption pairs in the evaluation set.

5.1.4 Image and Caption Similarity

Image and Caption Similarity is a relevance-based metric that measures the semantic alignment between a medical image and a generated caption. Unlike traditional text-only metrics, this approach directly evaluates cross-modal consistency by embedding both the image and caption into a shared representation space and computing their cosine similarity.

The implementation relies on a pretrained medical vision-language model introduced by Rückert et al. [Rüc+24b]. Given a medical image and a corresponding caption, the model independently encodes each modality and computes the cosine similarity between the resulting embeddings. The final score reflects how well the textual description semantically matches the visual content.

Formally, the similarity score is computed as:

$$\text{Score} = w \cdot \max\left(0, \cos(\mathbf{v}, \mathbf{c})\right) \tag{5.4}$$

where $\mathbf{v}$ and $\mathbf{c}$ are the image and caption embeddings, respectively, and $w = 2.5$ is a scaling factor used in the competition implementation. Cosine similarity is clipped at zero to avoid negative scores.

This metric has the advantage of capturing cross-modal semantic alignment, especially in a domain where visual and textual signals must correspond precisely. However, it is sensitive to the quality and domain-specific training of the underlying embedding model, which can affect generalizability to rare findings or non-standard phrasings.

5.1.5 AlignScore

AlignScore [Wan+23a] is a factuality-focused evaluation metric designed to measure the consistency of generated text with a reference. It was developed for applications such as

fact-checking, summarization, and medical report generation, where ensuring alignment between source and output is essential.

The metric works by comparing the generated caption (treated as a series of factual claims) with the reference caption (serving as context). Using a RoBERTa-based alignment model [Liu+19], AlignScore splits both the candidate and reference into semantically meaningful chunks. Each claim sentence from the candidate is then aligned to the most supportive context chunk from the reference, and a relevance score is computed for each pair.

The final AlignScore is the average of all claim–context alignment scores:

$$\text{AlignScore} = \frac{1}{N} \sum_{i=1}^{N} \max_{j} \text{align}(c_i, r_j) \tag{5.5}$$

where c_i are the claim chunks from the candidate caption, r_j are the reference chunks, $\text{align}(c_i, r_j)$ is the predicted alignment score between them, and N is the total number of candidate claim chunks.

AlignScore is particularly well-suited for tasks where factual accuracy matters more than surface similarity. Compared to semantic similarity metrics like BERTScore [Zha+20] or BLEURT [SDP20], AlignScore has been shown to better detect factual inconsistencies, especially in settings involving negation, contradiction, or omission [Wan+23a]. However, its performance depends on accurate chunking and alignment modeling, and it can be sensitive to sentence boundaries or fragmented input.

5.1.6 UMLS Concept F1

The UMLS Concept F1 score is a domain-specific evaluation metric that assesses the clinical accuracy of generated text by comparing its medical concept content with that of a reference. It leverages the Unified Medical Language System (UMLS) [Bod04] to extract standardized medical entities from both the candidate and the reference captions.

To perform this evaluation, both texts are first processed using a concept extraction tool—specifically, QuickUMLS[2] or MedCAT [Kra+19]—to identify UMLS concepts. Only entities that fall within specific semantic types (e.g., disorders, anatomy, procedures) are retained, consistent with the configuration used in the MEDCON metric [Yim+23]. The extracted concept sets are then compared using the the Dice–Sørensen coefficient[3]:

[2] https://pypi.org/project/quickumls/
[3] https://en.wikipedia.org/wiki/Dice-S%C3%B8rensen_coefficient

$$\text{UMLS-F1} = \frac{2 \cdot |C_{\text{cand}} \cap C_{\text{ref}}|}{|C_{\text{cand}}| + |C_{\text{ref}}|} \tag{5.6}$$

Here, C_{cand} and C_{ref} denote the sets of extracted UMLS concepts from the candidate and reference texts, respectively. The metric rewards systems that can identify and correctly include medically relevant content, independent of phrasing.

While this metric is highly suitable for clinical contexts, especially where factual correctness is paramount, it also comes with limitations. It is sensitive to named entity recognition performance and may penalize semantically correct but differently phrased outputs that use non-standard terminology not captured by UMLS.

5.2 Experimental Results

This section presents the evaluation of the DC systems developed in this thesis. The results are analyzed both qualitatively, by examining generated captions and model behaviors, and quantitatively, through metric-based comparisons on the development set. Additionally, the performance of selected models in the official test set of the ImageCLEFmedical 2025 Caption Prediction sub-task is reported.

5.2.1 Qualitative Evaluation

Benchmarking model performance alone does not capture the full range of model behavior, especially in complex, high-stakes domains such as medical image captioning. Qualitative analysis offers complementary insight by examining how captioning models behave under different conditions, the types of outputs they generate, and the nature of their errors or successes. This section presents selected examples that reveal important patterns and distinctions between models and post-processing techniques.

One such example involves the role of instruction phrasing in guiding model behavior. A key design feature of instruction-tuned models like InstructBLIP (Section 3.1) is their ability to adapt responses based on the textual prompt provided. In the context of diagnostic captioning, the specificity and clarity of the instruction can significantly affect the quality and relevance of the generated output. To explore this behavior, a controlled experiment was conducted in which a single medical image was paired with three instruction variants, ranging from no instruction to a highly descriptive prompt explicitly framing the task. The results of this experiment are presented in Figure 5.2. In the absence of any instruction, the model generates a generic and clinically uninformative description that merely notes the presence of a black and white image without identifying any anatomical structures

or medical findings. When provided with a slightly more specific prompt, asking it to describe the radiology image, the model produces a more relevant caption indicating that the chest appears normal with no visible abnormalities. However, it still avoids clinical terminology and does not refer to any specific features. In contrast, the most detailed prompt, which frames the task from the perspective of an experienced radiologist and requests a comprehensive description, leads to a much richer and more clinically grounded output. This version explicitly mentions the lungs, absence of pathological signs such as consolidation or pleural effusion, and confirms that the cardiac silhouette and mediastinum are within normal limits. These results demonstrate the significant impact that prompt phrasing can have on the semantic depth and clinical utility of the generated captions. More detailed instructions enable the model to produce outputs that are not only linguistically fluent but also closer to the expectations of professional medical reporting. At the same time, this highlights the importance of instruction clarity and design in guiding the behavior of instruction-tuned models, especially in high-stakes domains such as medical imaging.

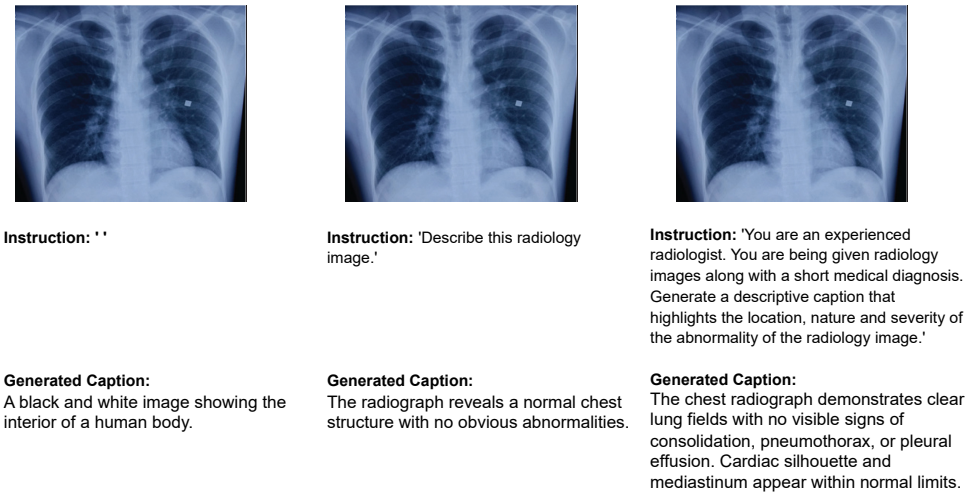


Fig. 5.2: Effect of instruction phrasing on captions generated by InstructBLIP. The same image is paired with three different instructions, resulting in varying levels of detail and clinical relevance.

Another aspect of qualitative evaluation focuses on the effectiveness of MedCLIP, described in Section 3.4, as a test-time reranking mechanism. In this experiment, MedCLIP is applied on top of the InstructBLIP checkpoint that achieved the highest score on the development set. A medical image from the development split is selected, and $m = 4$ diverse captions are generated using top-p sampling. These candidate captions are then scored by MedCLIP based on their visual-semantic alignment with the input image. The caption receiving the highest score is selected as the preferred output. For reference, the original ground truth caption is also included. Figure 5.3 presents the input image, the four sampled candidate captions along with their MedCLIP scores, and the corresponding ground truth. In this example, MedCLIP assigns the highest score to a caption describing a dislocated hip joint,

which diverges from the annotated reference indicating a superiorly located acetabular cyst. While the selected caption is syntactically fluent and anatomically reasonable, it does not capture the specific pathological finding of the reference caption. This outcome highlights both the benefits and limitations of embedding-based reranking. MedCLIP effectively promotes captions that are visually grounded and coherent with the image but may still favor more general or common radiological patterns over less frequent but clinically important findings. Moreover, inconsistencies in anatomical localization across captions (e.g., references to both the left and right hip) remain unpenalized, underscoring the need for more fine-grained factuality-aware mechanisms. Overall, this case illustrates that while MedCLIP improves visual-textual alignment, it may not be sufficient on its own for selecting clinically optimal captions.

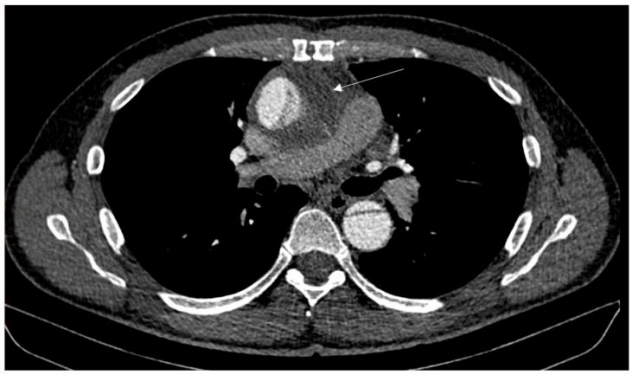
Ground Truth Caption:
Anteroposterior x-ray of right hip with a superiorly located acetabular cyst.

Caption 1: Radiograph of the right hip showing a lytic lesion with a left femoral neck fracture. **Score:** 0.1734
Caption 2: Postoperative X-ray of right hip showing a dislocated hip joint. **Score:** 0.2780
Caption 3: Postoperative AP pelvic X-ray showing bone graft implantation. **Score:** 0.2761
Caption 4: Radiograph of the left hip showing a lytic lesion involving the femoral head, acetabulum, and femoral neck. **Score:** 0.2725

Fig. 5.3: Caption reranking using MedCLIP. Among four sampled candidates, the caption with the highest alignment score is selected as the final output.

To further investigate model behavior, a final qualitative example is presented in which multiple captioning systems are prompted with the same medical image from the official test set. Figure 5.4 shows the input image alongside a table of captions generated by each system. The outputs illustrate notable variations: some models localize pathology differently (e.g., right upper lobe mass vs. bilateral effusions), others differ in the level of detail (specific mass description vs. general effusion), and one highlights imaging modality more precisely (angiography vs. CT). These differences suggest that while all systems generate coherent captions, they vary in clinical focus and emphasis, with some prioritizing common or salient findings over finer pathological details. All models were prompted

using the same instruction to ensure a fair and controlled comparison, as displayed in the table on the right side of Figure 5.2.



Model	Generated Caption
InstructBLIP (Instruction) [see Section 3.1]	Computed tomography scan of the chest showing a mass in the right upper lobe of the thorax.
InstructBLIP (Contrastive) [see Section 3.2]	Computed tomography (CT) scan of the chest showing bilateral pleural effusions.
Mixer [see Section 3.3]	Computed tomography (CT) scan of the chest showing a pleural effusion.
InstructBLIP (Instruction) + MedCLIP [see Section 3.4]	Computed tomography angiography of the chest showing a mass in the right upper lobe of the thorax.

Fig. 5.4: Input image and corresponding captions generated by different models. The models vary in architecture, tuning strategy, and inference behavior. The gold caption is: "Computed tomography scan of the chest showing a right hilar mass with associated mediastinal lymphadenopathy."

5.2.2 Quantitative Evaluation

Quantitative analysis provides a systematic evaluation of model performance using predefined metrics. As described in Chapter 4, the dataset was partitioned into three subsets: a training set of 73,027 samples, a validation set of 9,736 samples, and a development set of 14,605 samples. The training set was used to optimize models that required parameter updates, while the validation set served to monitor generalization performance during training. In particular, validation loss was used to determine early stopping points based on a predefined patience threshold. The development set, which remained untouched during training, was used as a private test set to evaluate all implemented models and their variations under consistent conditions.

The InstructBLIP model was trained for up to 40 epochs for the instruction-tuned version and 20 epochs for the contrastive version, using a batch size of 4. Training followed the detailed instruction prompt illustrated in Figure 5.2. Early stopping with a patience of 3 epochs was employed, terminating training at epochs 38 and 17 for the instruction-tuned and contrastive models, respectively, once validation loss showed no further improvement. The Mixer model was trained under two configurations. The first configuration

corresponds to the submission reported in Section 5.2.3, where training was limited to just 3 epochs due to time constraints associated with the ImageCLEFmedical 2025 deadline. The second configuration extended training to 12 epochs and aimed to fully leverage the reinforcement signal as described in Section 3.3. Due to the resource-intensive evaluation routine, particularly its memory demands, training was conducted with a batch size of 1. In addition, a stratified subset of the training and validation sets was used, selected based on the distribution of Concept Unique Identifiers (CUIs) provided in the Concept Detection task. Both configurations are reported to highlight the difference between undertrained and adequately trained models, with the latter gradually transitioning from optimizing for linguistic fluency to focusing on metric-based clinical accuracy as the reinforcement coefficient α increased.

All models were assessed using the six evaluation metrics described in Section 5.1. In addition to reporting individual metric scores, the results also include group-wise averages across relevance-based and factuality-based metrics, as well as an overall average defined as the mean of these two groups.

Model	Method	Overall	Similarity	BERTScore	ROUGE-1	BLEURT	Rel. Avg.	UMLS F1	AlignScore	Fact. Avg.
InstructBLIP (Instruction)	SFT	0.2977	0.7996	0.5919	**0.2161**	**0.3023**	0.4775	**0.1448**	0.0913	0.1180
InstructBLIP (Contrastive)	SFT	0.3031	**0.8354**	0.5904	0.1928	0.2939	**0.4781**	0.1442	0.1123	0.1282
Mixer (3 epochs)	RL	0.2959	0.6863	0.5461	0.1779	0.2544	0.4162	0.1059	**0.2453**	0.1756
Mixer (12 epochs)	RL	**0.3086**	0.7365	0.5591	0.1861	0.2618	0.4359	0.1250	0.2376	**0.1813**
InstructBLIP + MedCLIP	TTS	0.2986	0.7979	**0.5974**	0.2110	0.2899	0.4740	0.1436	0.1031	0.1233
Mixer (12 epochs) + MedCLIP	TTS	0.3078	0.7354	0.5667	0.1821	0.2596	0.4359	0.1271	0.2324	0.1797

Tab. 5.1: Evaluation results for all implemented methods on our held-out development set. Metrics are grouped into relevance-based (Similarity, BERTScore, ROUGE-1, BLEURT) and factuality-based (UMLS F1, AlignScore), with average scores computed per group and an overall average.

Table 5.1 summarizes the quantitative performance of all implemented models, with each system additionally tagged by its post-training method (SFT, RL, or TTS). Several patterns emerge. First, the two InstructBLIP variants perform strongly on relevance-based metrics, namely Similarity, BERTScore, ROUGE-1, and BLEURT, with the contrastive version achieving the highest similarity score (0.8354), while the instruction-tuned model leads on ROUGE-1 (0.2161) and BLEURT (0.3023). However, both variants show comparatively low scores on the factuality-based AlignScore metric (0.0913 and 0.1123), suggesting that they prioritize fluent and semantically close captions but often overlook finer clinical correctness. In contrast, the Mixer models stand out on factuality-oriented evaluation: despite lower similarity scores (0.6863 and 0.7365 compared to > 0.79 for InstructBLIP), the 12-epoch Mixer and its MedCLIP variant achieve some of the highest AlignScore values (0.2376 and 0.2324), reflecting more accurate alignment with clinical entities. Interestingly, extending Mixer training from 3 to 12 epochs yields clear improvements in both overall score (from 0.2959 to 0.3086) and factuality average (from 0.1756 to 0.1813), underscoring the role of adequate training time. TThe addition of MedCLIP has a modest but mixed effect on factuality: it slightly increases UMLS F1 (from 0.1250 to 0.1271) but reduces AlignScore (from 0.2376 to 0.2324), without major gains in relevance. Overall, these

results highlight a trade-off between relevance and factual accuracy: InstructBLIP excels in producing linguistically and semantically coherent captions (e.g., Similarity > 0.79 across both variants), whereas Mixer models, especially when sufficiently trained, better capture clinically meaningful information (AlignScore > 0.23).

5.2.3 ImageCLEFmedical Caption 2025 Submissions

A key part of this thesis was developed in the context of the AUEB NLP Group's participation in the ImageCLEFmedical 2025 Caption Prediction sub-task. We submitted a series of systems based on the methods described in Sections 3.1, 3.3, and 3.4, either as standalone models or in various combinations. These models incorporate instruction tuning (InstructBLIP), reinforcement-driven optimization (Mixer), and caption reranking (MedCLIP), aiming to improve both linguistic quality and clinical fidelity in diagnostic caption generation.

Our best-performing submission ranked 5$^{\text{th}}$ out of 8 participating teams, demonstrating competitive performance across a wide range of evaluation criteria. Table 5.2 summarizes the metric scores achieved by our systems on the competition's official test set. These scores differ from those reported elsewhere in this thesis, as the competition used a distinct held-out test set that was not publicly accessible.

When interpreting these results, it is important to note that the MedCLIP component was used strictly as a post-processing reranker rather than being integrated into training, which partly explains the modest gains observed. Moreover, the Mixer-based systems were trained for only three epochs due to time constraints, meaning that their reported scores are not fully representative of the method's potential. As shown in our quantitative experiments in Section 5.2.2, the same approach achieves substantially stronger results on the development set when trained for a sufficient number of epochs. This suggests that the relative competitiveness of Mixer is likely underestimated in the official leaderboard.

ID	System	Overall	Similarity	BERTScore	ROUGE-1	BLEURT	Rel. Avg.	UMLS F1	AlignScore	Fact. Avg.	Rank
1403	InstructBLIP	**0.3068**	**0.7947**	0.5884	**0.2176**	**0.3030**	**0.4759**	**0.1429**	0.1325	0.1377	**48**
1724	InstructBLIP + MedCLIP Reranker	0.3026	0.7896	**0.5939**	0.2122	0.2897	0.4714	0.1421	0.1257	0.1339	61
1960	Mixer (Sampling)	0.2853	0.6778	0.5453	0.1814	0.2583	0.4157	0.1038	**0.2058**	**0.1548**	83
1962	Mixer (Sampling) + MedCLIP Reranker	0.2757	0.6539	0.5621	0.1868	0.2585	0.4153	0.1037	0.1684	0.1361	88
1961	Mixer (Beam)	0.2747	0.6649	0.5472	0.1814	0.2637	0.4143	0.0998	0.1706	0.1352	89
1963	Mixer (Beam) + MedCLIP Reranker	0.2732	0.6498	0.5560	0.1886	0.2579	0.4140	0.1022	0.1627	0.1324	91

Tab. 5.2: Evaluation metrics for all submissions to the ImageCLEFmedical 2025 Caption Prediction task. Metrics are grouped into Relevance (Similarity, BERTScore, ROUGE-1, BLEURT) and Factuality (UMLS F1, AlignScore). The Relevance and Factuality columns report the average score within each group, and the Overall score is the mean of those two averages. Including system names directly with IDs avoids the need for a separate mapping table.

Table 5.1 provides additional context for interpreting the official ImageCLEFmedical 2025 results shown in Table 5.2. The trends observed on our held-out development set mirror those on the competition leaderboard: InstructBLIP variants excel on relevance-based met-

rics, while Mixer models, especially when trained for 12 epochs, achieve higher factuality scores. This pattern helps explain why the official submissions with Mixer trained for only three epochs appear less competitive; given sufficient training, Mixer's factual accuracy, and overall performance, would likely improve further. The modest impact of MedCLIP observed in both tables also confirms its role as a post-processing reranker: it provides small but consistent gains in factuality without substantially affecting relevance. Together, these results highlight the trade-off between linguistic relevance and clinical fidelity and demonstrate how development-set experiments can illuminate the relative strengths and limitations of each method in the official leaderboard.

6 Conclusions and Future Work

6.1 Conclusions

This thesis examined post-training strategies for adapting general-purpose vision-language models to the task of diagnostic captioning. The motivation stems from the domain-specific requirements of medical imaging, where clinical accuracy, factual consistency, and interpretability are essential. To this end, the work focused on three post-training approaches: Supervised Fine-Tuning (SFT), which included both standard cross-entropy training and a variant augmented with a contrastive loss to improve alignment between image and text representations, Reinforcement Learning (RL) with metric-driven optimization, and Test-Time Scaling (TTS) through caption reranking mechanisms. A series of methods were implemented and evaluated on the ImageCLEFmedical 2025 dataset, with a primary focus on quantitative metrics and a limited qualitative assessment. The experimental results revealed the complementary strengths of different strategies. Instruction tuning produced fluent and adaptable outputs when guided by well-crafted prompts. RL contributed to improved alignment with evaluation metrics, particularly those measuring factual correctness, as shown in Table 5.1. Caption reranking with a domain-specific vision–language model helped prioritize outputs that were more consistent with the visual content, improving factual grounding even when the selected caption did not fully match the ground truth description, as shown in Figure 5.3. The findings highlight the potential of integrating multiple post-training strategies to address different aspects of model behavior, offering a comprehensive framework for developing diagnostic captioning systems. Participation in the ImageCLEFmedical 2025 challenge further validated the developed systems, with competitive rankings demonstrating their effectiveness in real-world evaluation settings. Overall, this work reinforces the value of targeted adaptation in high-stakes applications of AI. By bridging the gap between generic pretrained models and specialized clinical requirements, it lays the foundation for further research into robust, explainable, and trustworthy medical captioning systems.

6.2 Future Work

While this thesis has demonstrated the potential of post-training techniques for improving diagnostic captioning performance, several promising directions remain for future explo-

ration. One such direction involves incorporating Curriculum Reinforcement Learning, where training is structured to progress from simpler to more complex diagnostic examples. In the context of the ImageCLEFmedical dataset, case difficulty could be estimated using factors such as caption length, the number of UMLS concepts mentioned, or the rarity of those concepts across the dataset. Simpler cases could be prioritized early in training, while more complex or rare ones could be introduced later or assigned greater weight in the reward computation. Such a curriculum could be integrated into the existing reward function used in the RL-based Mixer model, enabling the system to gradually improve its clinical reasoning and robustness.

Another promising avenue lies in extending TTS with search-based methods such as Monte Carlo Tree Search (MCTS). Unlike the MedCLIP-based caption reranking, which evaluates a fixed set of candidate captions post hoc to select the most visually grounded one, MCTS actively guides the generation process by exploring multiple candidate sequences in parallel during decoding. At each step, MCTS evaluates partially generated captions using a reward function, potentially incorporating metrics for relevance, factuality, or clinical accuracy, allowing the search to prioritize paths that are likely to produce high-quality final outputs. In this way, MCTS is not merely a replacement of the scoring function after generation, but a dynamic decoding strategy that iteratively informs which tokens to generate next, offering a more fine-grained control over fluency and clinical correctness compared to standard greedy, beam, or post-hoc reranking approaches.

Lastly, the continued evolution of large-scale multimodal language models opens the door to further improvements. Recent models like MedGemma[1], developed specifically for biomedical visual-language tasks, offer domain-specialized knowledge that could be highly beneficial for diagnostic captioning. Future work could investigate their capabilities either as standalone models or as initialization points for fine-tuning. Additionally, evaluating and adapting newly introduced general-purpose LLMs to the medical setting remains an exciting research opportunity, particularly as instruction tuning and zero-shot prompting continue to evolve. Overall, these directions highlight the potential for building diagnostic captioning systems that are more accurate, adaptive, and aligned with the demands of real-world clinical environments.

[1] https://medgemma.org/

Bibliography

[Ala+22] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. "Flamingo: A Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[And+18] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6077–6086. DOI: 10.1109/cvpr.2018.00636.

[BAM17] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2017), pp. 423–443. DOI: 10.1109/tpami.2018.2798607.

[Bod04] O. Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D267–D270. DOI: 10.1093/nar/gkh061.

[Boj+17] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. "Findings of the 2017 Conference on Machine Translation (WMT17)". In: *Proceedings of the Second Conference on Machine Translation (WMT)*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 169–214.

[Bro+20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*. NeurIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[Caf+24] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara. "The Revolution of Multimodal Large Language Models: A Survey". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2024, pp. 13590–13618. DOI: 10.18653/v1/2024.findings-acl.807.

[Cha+21] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption Tasks 2021". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21–24*. Vol. 2936. CEUR Workshop Proceedings. 2021, pp. 1184–1200.

[Cha+22] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2022". In: *CLEF 2022 Working Notes*. CEUR Workshop Proceedings. Bologna, Italy: CEUR-WS.org, 2022, pp. 1355–1373.

[Cha+25] A. Chatzipapadopoulou, I. Pantelidis, F. Charalampakos, M. Samprovalaki, G. Moschovis, P. Kaliosis, K. V. Dalakleidi, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2025". In: *CLEF 2025 Working Notes*. Ed. by G. Faggioli, N. Ferro, P. Rosso, and D. Spina. Madrid, Spain, 2025, pp. 2384–2407.

[Cha25] A. Chatzipapadopoulou. "Enhanced Biomedical Image Tagging". Bachelor's Thesis. Athens University of Economics and Business, Department of Informatics, 2025. URL: http://nlp.cs.aueb.gr/theses/Bsc_Thesis_Chatzipapadopoulou.pdf.

[Chu+24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. "Scaling Instruction-Finetuned Language Models". In: *Journal of Machine Learning Research* 25 (2024), 70:1–70:53.

[Dai+23] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi. "InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[Dam+25] 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. Ben Abacha, A. García Seco de Herrera, H. Müller, and C. M. Friedrich. "Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation". In: *CLEF 2025 Working Notes*. CEUR Workshop Proceedings. Madrid, Spain: CEUR-WS.org, Sept. 2025.

[Dem+16] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. "Preparing a Collection of Radiology Examinations for Distribution and Retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (Mar. 2016), pp. 304–310. ISSN: 1067-5027. DOI: 10.1093/jamia/ocv080.

[Dev+19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[Gao+19] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao. "Self-Critical N-Step Training for Image Captioning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6300–6308. DOI: 10.1109/cvpr.2019.00646.

[Han+20] Z. Han, B. Wei, Y. Yin, and S. Li. "Unifying Neural Learning and Symbolic Reasoning for Spinal Medical Report Generation". In: *Medical Image Analysis* 67 (2020), p. 101872. DOI: 10.1016/j.media.2020.101872.

[HB21] S. Hanna and O. Bojar. "A Fine-Grained Analysis of BERTScore". In: *Proceedings of the Sixth Conference on Machine Translation (WMT)*. Online: Association for Computational Linguistics, 2021, pp. 880–895. URL: https://aclanthology.org/2021.wmt-1.59.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

[Hos+18] M. Z. Hossain, F. Sohel, M. Shiratuddin, and H. Laga. "A Comprehensive Survey of Deep Learning for Image Captioning". In: *ACM Computing Surveys* 51 (2018), pp. 1–36. DOI: 10.1145/3295748.

[HS97] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[Ion+25] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. 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, and 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). Madrid, Spain: Springer Lecture Notes in Computer Science LNCS, Sept. 2025.

[Joh+19] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, S. Horng, L. Ngo, D. J. Stone, et al. "MIMIC-CXR, a Large Publicly Available Database of Labeled Chest Radiographs". In: *Scientific Data* 6.1 (2019), pp. 1–8.

[JXX18] B. Jing, P. Xie, and E. Xing. "On the Automatic Generation of Medical Imaging Reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018, pp. 2577–2586. DOI: 10.18653/v1/p18-1240.

[Kal+23] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2023". In: *CLEF 2023 Working Notes*. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, 2023.

[Kal+24] P. Kaliosis, J. Pavlopoulos, F. Charalampakos, G. Moschovis, and I. Androutsopoulos. "A Data-Driven Guided Decoding Mechanism for Diagnostic Captioning". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 7450–7466. URL: https://aclanthology.org/2024.findings-acl.444.

[Kas+23] Ö. Kasalak, H. Alnahwi, R. Toxopeus, J. P. Pennings, D. Yakar, and T. C. Kwee. "Work overload and diagnostic errors in radiology". In: *European Journal of Radiology* 167 (2023), p. 111032. ISSN: 0720-048X. DOI: https://doi.org/10.1016/j.ejrad.2023.111032. URL: https://www.sciencedirect.com/science/article/pii/S0720048X23003467.

[Kes+19] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher. *CTRL: A Conditional Transformer Language Model for Controllable Generation*. 2019. arXiv: 1909.05858 [cs.CL]. URL: https://arxiv.org/abs/1909.05858.

[Koj+22] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. "Large Language Models are Zero-Shot Reasoners". In: *arXiv preprint arXiv:2205.11916* (2022).

[KP25] P. Kaliosis and J. Pavlopoulos. "Learning to Align: Addressing Character Frequency Distribution Shifts in Handwritten Text Recognition". In: *arXiv preprint arXiv:2506.09846* (2025). URL: https://arxiv.org/abs/2506.09846.

[Kra+19] Z. Kraljević, D. Bean, A. Mascio, Ł. Roguski, A. Folarin, A. Roberts, R. Bendayan, and R. Dobson. "MedCAT: Medical Concept Annotation Tool". In: *arXiv preprint arXiv:1912.10166* (2019). URL: https://arxiv.org/abs/1912.10166.

[Li+20] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. "Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 121–137. DOI: 10.1007/978-3-030-58577-8_8.

[Li+22] J. Li, D. Li, C. Xiong, and S. Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. 2022, pp. 12888–12900.

[Li+23] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023, pp. 19730–19742.

[Lin+14] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft COCO: Common Objects in Context". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[Lin04] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[Liu+19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019). URL: https://arxiv.org/abs/1907.11692.

[Liu+21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9992–10002. DOI: 10.1109/iccv48922.2021.00986.

[Moo+23] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. "Med-Flamingo: A Multimodal Medical Few-Shot Learner". In: *Proceedings of the Machine Learning for Health Workshop (ML4H) at NeurIPS*. 2023, pp. 353–367.

[NDK23] A. Nicolson, J. Dowling, and B. Koopman. "A Concise Model for Medical Image Captioning". In: *Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*. Thessaloniki, Greece, Sept. 2023.

[OLV18] A. van den Oord, Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *arXiv preprint arXiv:1807.03748* (2018). URL: http://arxiv.org/abs/1807.03748.

[Ouy+22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. J. Lowe. "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[Pap+02] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002, pp. 311–318.

[Pav+21] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail. "Diagnostic Captioning: A Survey". In: *Knowledge and Information Systems* 64 (2021), pp. 1691–1722. DOI: 10.1007/s10115-022-01684-7.

[Rad+21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[Raf+19] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21 (2019), 140:1–140:67.

[REM23] R. Ramos, D. Elliott, and B. Martins. "Retrieval-Augmented Image Captioning". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2023, pp. 3666–3681. DOI: 10.18653/v1/2023.eacl-main.266.

[Ren+15] S. Ren, K. He, R. B. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), pp. 1137–1149. DOI: 10.1109/tpami.2016.2577031.

[Ren+17a] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1151–1159. DOI: 10.1109/cvpr.2017.128.

[Ren+17b] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. "Self-Critical Sequence Training for Image Captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7008–7024.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[Rüc+24a] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. Horn, F. Nensa, and C. M. Friedrich. "ROCOv2: Radiology Objects in Context Version 2, an Updated Multimodal Image Dataset". In: *Scientific Data* 11.1 (2024). DOI: 10.1038/s41597-024-03496-6.

[Rüc+24b] J. Rückert, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, and C. M. Friedrich. "MedImageInsights: A Pretrained Model for Interpretable Medical Image Captioning". In: *arXiv preprint arXiv:2410.06542* (2024).

[Saa+24] K. Saab, T. Tu, W.-H. Weng, R. Tanno, et al. "Capabilities of Gemini Models in Medicine". In: *arXiv preprint arXiv:2404.18416* (2024).

[Sam+24] M. Samprovalaki, A. Chatzipapadopoulou, G. Moschovis, F. Charalampakos, P. Kaliosis, J. Pavlopoulos, and I. Androutsopoulos. *AUEB NLP Group at ImageCLEFmedical 2024*. Grenoble, France, 2024.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd. MIT Press, 2018.

[Sch+22] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[SD25] K. Sun and M. Dredze. "Amuro & Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models". In: *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*. 2025, pp. 131–151. DOI: 10.18653/v1/2025.repl4nlp-1.11.

[SDP20] T. Sellam, D. Das, and A. P. Parikh. "BLEURT: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2020, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704.

[Sha+18] P. Sharma, N. Ding, S. Goodman, and R. Soricut. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018, pp. 2556–2565.

[SHB16] R. Sennrich, B. Haddow, and A. Birch. "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*. Association for Computational Linguistics, 2016, pp. 86–96. DOI: 10.18653/v1/P16-1009. URL: https://aclanthology.org/P16-1009/.

[Sîr+25] I. Sîrbu, I.-R. Sîrbu, J. Bogojeska, and T. Rebedea. "GIT-CXR: End-to-End Transformer for Chest X-Ray Report Generation". In: *Information* 16.7 (2025), p. 524. DOI: 10.3390/info16070524.

[Sze+15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/cvpr.2015.7298594.

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008.

[Vin+15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and Tell: A Neural Image Caption Generator". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3156–3164. DOI: 10.1109/cvpr.2015.7298935.

[Wan+22] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text". In: *arXiv preprint arXiv:2210.10163* (2022).

[Wan+23a] Y. Wan, D. Tam, M. Bansal, and W.-T. Yih. "AlignScore: Evaluating Factual Consistency with A Unified Alignment Function". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2023, pp. 11386–11403. DOI: 10.18653/v1/2023.acl-long.634.

[Wan+23b] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. "Self-Instruct: Aligning Language Models with Self-Generated Instructions". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 2023, pp. 13484–13508. DOI: 10.18653/v1/2023.acl-long.754.

[Wei+22a] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. "Finetuned Language Models Are Zero-Shot Learners". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2022.

[Wei+22b] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[Wei+23] J. Wei, X. Wang, D. Schuurmans, E. H. Chi, Q. V. Le, and D. Zhou. "LLM Post-Training: A Deep Dive into Reasoning". In: *arXiv preprint arXiv:2309.16747* (2023). URL: https://arxiv.org/abs/2309.16747.

[Xu+15] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015, pp. 2048–2057.

[Yan+23] S. Yang, X. Wu, S. Ge, X. Wu, S. K. Zhou, and L. Xiao. "Radiology Report Generation with a Learned Knowledge Base and Multi-Modal Alignment". In: *Medical Image Analysis* 86 (2023), p. 102798. DOI: 10.1016/j.media.2023.102798.

[Yim+23] W.-W. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. Yetisgen. "Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation". In: *Scientific Data* 10.1 (2023), p. 586. DOI: 10.1038/s41597-023-02487-3.

[YX+24] L. Yang, S. Xu, et al. "Advancing Multimodal Medical Capabilities of Gemini". In: *arXiv preprint arXiv:2405.03162* (2024).

[Zha+20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations (ICLR)*. 2020.

[Zha+21] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. "VinVL: Revisiting Visual Representations in Vision-Language Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 5579–5588. DOI: 10.1109/cvpr46437.2021.00553.

[Zho+20] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao. "Unified Vision-Language Pre-Training for Image Captioning and VQA". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020. DOI: 10.1609/aaai.v34i07.7005.

List of Acronyms

CV	Computer Vision
NLP	Natural Language Processing
DC	Diagnostic Captioning
SFT	Supervised Fine-Tuning
RL	Reinforcement Learning
TTS	Test-Time Scaling
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
DL	Deep Learning
VLMs	Vision-Language Models
LLMs	Large Language Models
AI	Artificial Intelligence
SSL	Self-Supervised Learning
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
CL	Contrastive Learning

CLIP	Contrastive Language–Image Pre-training
ITM	Image-Text Matching
MIM	Masked Image Modeling
MMLM	Multimodal Masked Language Modeling
ViT	Vision Transformer
RLHF	Reinforcement Learning with Human Feedback
CoT	Chain-of-Thought
RAG	Retrieval-Augmented Generation
DMMCS	Distance from Median Maximum Concept Similarity
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BLIP	Bootstrapping Language-Image Pre-training
ITC	Image-Text Contrastive
LM	Language Modeling
MLLMs	Multimodal Large Language Models
InfoNCE	Information Noise-Contrastive Estimation
SCST	Self-Critical Sequence Training
ROCOv2	Radiology Objects in COntext Version 2
UMLS	Unified Medical Language System
IDF	Inverse Document Frequency

ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BLEURT	Bilingual Evaluation Understudy with Representations from Transformers
CUIs	Concept Unique Identifiers
MCTS	Monte Carlo Tree Search

List of Figures

List of Tables