

Enhancing Medical Image Captioning Through Multimodal Alignment: A Lightweight Approach

Magda T. Amorim

Master in Engineering Physics

Department of Physics and Astronomy

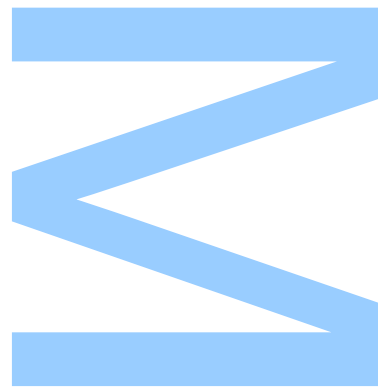
Faculty of Sciences and Faculty of Engineering of the University of Porto

2025

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



Enhancing Medical Image Captioning Through Multimodal Alignment: A Lightweight Approach

Magda T. Amorim

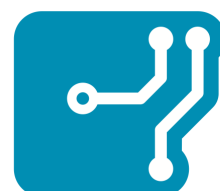
Dissertation carried out as part of the Master in Engineering
Physics
Department of Physics and Astronomy
2025

Supervisor

Luís F. Teixeira, Faculty of Engineering

Co-Supervisor

Hélder P. Oliveira, Faculty of Sciences



INESCTEC
| ASSOCIATE LABORATORY
| PORTUGAL

UNIVERSITY OF PORTO

MASTERS THESIS

Enhancing Medical Image Captioning Through Multimodal Alignment: A Lightweight Approach

Author:

Magda T. AMORIM

Supervisor:

Luís F. TEIXEIRA

Co-supervisor:

Helder P. OLIVEIRA

*A thesis submitted in fulfilment of the requirements
for the degree of Masters in Physics Engineering*

at the

Faculty of Engineering and Science of University of Porto

Department of Astronomy and Physics

INESC TEC Visual Computing and Machine Intelligence Group

August 8, 2025

“ The question of whether a computer can think is no more interesting than the question of whether a submarine can swim. ”

Alan M Turing

This thesis is dedicated to my beloved grandparents, Rosa and António, who taught me the meaning of perseverance; to my wonderful parents, Anabela and Carlos, for their endless belief in me; and to my dear sister, Matilde, whose presence made every step lighter. Your unwavering support, encouragement, and love have guided me throughout every step of my academic journey and personal achievements. I am deeply grateful for everything you have done.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Luís F. Teixeira, and co-supervisor, Dr. Hélder P. Oliveira, for their guidance. I am especially thankful to Dr. Hugo S. Oliveira for his outstanding knowledge sharing and continued dedication and unwavering support throughout this study. His insightful ideas and constant encouragement were instrumental in the successful completion of this thesis.

I also acknowledge the Faculty of Sciences and Engineering of the University of Porto for its role in shaping my academic journey and for the continuous support provided throughout my studies. A very special thanks to the INESC TEC research laboratory, mainly the Visual Computing and Machine Intelligence Group Team (VCMI) for the support and to Dr. Hugo S. Oliveira for providing the necessary resources and environment to carry out the experimental work.

I am deeply grateful to my sister and her network for their expert support and invaluable medical insights that greatly contributed to this study. My heartfelt thanks go to my parents for their unwavering support and patience throughout my academic journey.

Funding

This work was co-financed by the Fundação para a Ciência e a Tecnologia (FCT) during the initial stages through a Research Initiation Grant, and by INESC TEC and Dr. Hugo S. Oliveira, who provided essential computational resources for the success of this research.

University of Porto

Abstract

Faculty of Engineering and Science of University of Porto

Department of Astronomy and Physics

INESC TEC Visual Computing and Machine Intelligence Group

Masters in Physics Engineering

Enhancing Medical Image Captioning Through Multimodal Alignment: A Lightweight Approach

by [Magda T. AMORIM](#)

This thesis explores advanced methods to enhance medical image captioning for X-ray images where clinically significant labels guide the generation of precise and informative descriptions. Building on state-of-the-art baseline models, this study systematically evaluates various image and text pretrained models to optimize the clinical relevance and linguistic quality of generated captions. Medical domain specific metrics are employed to ensure rigorous and meaningful evaluation. Furthermore, the research investigates efficient fusion strategies that incorporate extracted clinical labels into the captioning pipeline without sacrificing model performance. The proposed framework, aiming to facilitate practical applications such as virtual medical assistants and automated radiology report generation, demonstrate significant improvements in caption accuracy and informativeness, advancing automated radiology reporting and supporting more informed clinical decision-making.

Keywords - Medical Image Captioning, Label-Conditioned Captioning, Multimodal Fusion, X-ray Imaging, Radiology Report Generation, Image Encoder, Text Decoder, Fusion Strategies, Domain-Specific Evaluation Metrics.

University of Porto

Resumo

Faculty of Engineering and Science of University of Porto

Department of Astronomy and Physics

INESC TEC Visual Computing and Machine Intelligence Group

Mestrado em Engenharia Física

Aprimorando a Legendagem de Imagens Médicas Por Meio do Alinhamento

Multimodal: Uma Abordagem Leve

por [Magda T. AMORIM](#)

Esta tese explora métodos avançados para aprimorar a legendagem de imagens médicas para imagens de raios-X, onde classificações clinicamente significativas guiam a geração de descrições precisas e informativas. Com base em modelos de última geração, este estudo avalia sistematicamente diversos codificadores de imagem e decodificadores de texto para otimizar a relevância clínica e a qualidade linguística das legendas geradas. Para garantir uma avaliação rigorosa e significativa, são utilizadas métricas específicas, adaptadas para imagens médicas. Além disso, a pesquisa investiga estratégias de fusão eficientes que incorporam classificações clínicas ao modelo de legendagem sem comprometer o desempenho do mesmo. A estrutura proposta visa facilitar aplicações práticas, como assistentes médicos virtuais e geração automática de relatórios de radiologia. Os resultados demonstram melhorias significativas na precisão e no conteúdo informativo das legendas, promovendo relatórios de radiologia automatizados e apoiando uma tomada de decisão clínica mais informada.

Palavras-chave - Legendas de Imagens Médicas, Legendas Condicionadas por Classificações, Fusão Multimodal, Imagem de Raios-X, Geração de Relatórios de Radiologia, Codificador de Imagens, Decodificador de Texto, Estratégias de Fusão, Métricas de Avaliação Específicas.

Contents

Acknowledgements	vii
Abstract	x
Resumo	xi
Contents	xiii
List of Figures	xvii
List of Tables	xxi
1 Introduction	1
1.1 Introduction	1
1.2 Motivation and Contextualization	2
1.3 Methodology Employed	3
1.4 Objectives	4
1.5 Outline of the Thesis	5
2 Literature Review	7
2.1 Multimodal	7
2.1.1 Multimodal Characteristics	7
2.1.1.1 Multimodal Fusion Approaches	8
2.1.2 Multimodal Image Captioning Approaches	11
2.2 Literature Search	13
2.2.1 Search Methodology	13
2.2.2 Review of Existing Surveys	15
2.2.3 Multimodal Datasets	16
2.2.4 Image Captioning Works	17
2.3 Conclusions	22
3 Methodologies and Procedures	23
3.1 Main Methodology	23
3.1.1 Image Captioning Workflow	24
3.1.1.1 Image Encoders	24
3.1.1.2 Text Encoders and Decoders	27
3.1.1.3 CLIP	29

3.2	Datasets	30
3.2.1	MIMIC-CXR Dataset	30
3.2.2	ROCO Dataset	33
3.2.3	NIH Dataset	34
3.3	Evaluation and Performance Metrics	35
3.3.1	Domain Specific Metrics	36
3.3.2	Traditional Natural Language Preprocessing (NLP) Metrics	36
3.3.3	Overview	39
3.4	Baseline Models	39
3.4.1	ClipCap Model	39
3.4.2	BLIP Model	40
3.4.3	MediCap Model	41
3.4.4	DiCO Model	42
3.5	Label Conditioned Image Captioning	42
3.6	Training Methods	46
3.7	Text Generation Methods	47
3.8	Conclusions	48
4	Experiments and Results	49
4.1	Performance Criteria and Model Selection	49
4.2	Baseline Models	50
4.2.1	Zero shot Results	51
4.2.2	Few shot Predictions	53
4.3	Encoder and Decoder Selection	55
4.3.1	Visual Explanations with Grad-CAM	59
4.4	Dataset Selection	60
4.5	Adaptations	63
4.5.1	First Adaptations	64
4.5.2	Second Adaptations	69
4.5.2.1	Gradients Analysis	73
4.5.2.2	Attention Analysis	73
4.6	Label Conditioned Image Captioning	74
4.7	Conclusions	79
5	Conclusions	81
5.1	Contributions	81
5.2	Future Work	83
A	Baseline Models	85
A.1	Zero shot Predictions	85
A.2	Few shot Predictions	86
B	Pretrained Submodels Selection	89
C	Dataset Selection and Adaptations	91
C.0.1	Dataset Selection	91
C.0.2	Adaptations	92

D Label Conditioned Captioning	95
D.0.1 Caption Multilabel Classifier	95
D.0.2 Label Conditioned Image Captioning	96
 Bibliography	 99

List of Figures

1.1	Diagram with three main areas in which Artificial Intelligence (AI) technology in medical field is currently being implemented: medical image analysis, omics analysis, and natural language processing (Hamamoto).	2
1.2	Diagram with the general Generative Based Image Caption Generation Pipeline with Mapping.	3
2.1	Core research challenges in multimodal learning (Liang et al., 2023).	8
2.2	Schematic of the common subspace learning, which aims to project the heterogeneous data of different modalities into a common subspace, where the multimodal data with similar semantics will be represented by a similar vector (Guo et al.).	9
2.3	Multimodal Fusion Techniques.	10
2.4	Image Captioning Methods.	11
2.5	Generative Approach: Full Training (left) and Mapping (right).	13
2.6	Search Methodology: Top-down	13
2.7	Litmap: Foundational Models.	14
2.8	Trend of the dataset usage	17
2.9	A unified encoder-decoder model for general vision-language pre-training. The pre-trained model is then fine-tuned for image captioning and visual question answering (Zhou et al.)	19
2.10	Overview of proposed framework involving three components: encoder, decoder, and reward module. The encoder comprises two branches that extract visual and semantic features for the decoder separately. The decoder generates a report hierarchically. Note that the decoding process is not unfolded due to space limitations. In adversarial training, the generator is trained by Reinforcement Learning (RL) using rewards offered by the discriminators. Simultaneously, the discriminators are updated by the maximum-likelihood estimation approach (Hou et al.).	20
2.11	Overview of MMed-RAG, a versatile factual multimodal RAG system designed to enhance the reliability of Med-LVLMs. It introduces a domain-aware retrieval mechanism that effectively handles different domains of medical images by selecting suitable retrieval models. Additionally, it uses an adaptive context selection approach to determine the optimal number of retrieved contexts. It employs preference fine-tuning to improve both cross-modality and overall alignment (Xia et al.).	21
3.1	Standard Image Captioning Workflow	24
3.2	Architecture of LeNet-5 a Convolutional Neural Network for handwritten letters recognition Lecun et al..	25

3.3	Computation of one step of the convolution Roboflow.	25
3.4	Attention Mechanism Vaswani	26
3.5	Vision Transformer Dosovitskiy et al.	27
3.6	Recurrent Neural Network Architecture neural network.	27
3.7	Text Transformer Architecture Vaswani.	28
3.8	Overall pre-training and fine-tuning procedures for Bidirectional Encoder Representations from Transformers (BERT). Apart from output layers, the same architectures are used in pre-training and fine-tuning. The same pre-trained model parameters are used to initialise models for different downstream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token. Devlin et al.	29
3.9	Contrastive Language-Image Pre-Training (CLIP) jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples(Radford et al.).	30
3.10	Images which highlight the amount of variation present in MIMIC-CXR. . .	31
3.11	The steps required to extract Natural Language Explanation (NLE)s from raw radiology reports. Findings and Impressions are extracted first, then labels refer to each sentence, generating sentences that contain explanation keywords, Kayser et al..	31
3.12	The model pipeline to provide an NLE for a prediction Kayser et al.. . . .	32
3.13	Frequency of labels in MIMIC-CXR NLE on the training subset of 222,750 unique radiologic studies (8 studies were not labelled).	32
3.14	Examples of Radiology Objects in COContext (ROCO) images and respective captions: Left - "Chest radiograph. Parenchymal density in the right upper lobe." Right - "Detailed view of a single image in the teaching file. Using the navigation bar, all images of a case can be viewed sequentially in high resolution."	33
3.15	Overview of the workflow used for the ROCO dataset development Pelka et al..	34
3.16	Examples of NIH images and respective captions: Left - "XXXX-year-old XXXX with pain. Cardiomeastinal silhouette demonstrates normal heart size with tortuosity and atherosclerosis of the thoracic aorta. No focal consolidation, pneumothorax, or pleural effusion. No acute bony abnormality identified. Multilevel degenerative disc disease of the thoracic spine was noted. No acute cardiopulmonary abnormality."; Right -"Chest x-XXXX dated XXXX at XXXX hours. XXXX-year-old female, transplant workup. Calcified mediastinal XXXX. No focal areas of consolidation. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Degenerative changes thoracic spine. No acute cardiopulmonary abnormality."	34

3.17	a) Experimental design for selecting radiology reports and comparing metrics and radiologists in evaluating reports. b) Given a test report, select the report with the highest metric score from the training report corpus with respect to the test report and a particular metric. c) Conducting radiologist evaluation on the high metric score report relative to the test report, where radiologists identify the number of clinically significant and insignificant errors in the high metric score report across six error categories. d) Determining the alignment between metric scores and radiologists scores assigned to the same reports using the Kendall rank correlation coefficient. Yu et al.	35
3.18	Illustration of the computation of the recall metric RBERT. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red and include the optional importance weighting Zhang et al.. . . .	38
3.19	Overview of the transformer-based architecture, enabling the generation of meaningful captions while both CLIP and the language model, GPT-2, are frozen. We train a lightweight transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2 to extract a fixed-length prefix. It was employed GPT-2 at inference to generate the caption given the prefix embeddings. Also, suggest an MLP-based architecture (Mokady et al.).	39
3.20	Pre-training model architecture and objectives of BLIP (same parameters have the same color). Propose a multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions and is trained with an image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images (Li et al.)	40
3.21	Decoder conditioned on the visual features of the image via A) the cross-attention, and B) the self-attention. The visual features are extracted with the encoder. N is the number of Transformer blocks. [BOS] is the special token for the beginning of the sentence (Nicolson et al.).	41
3.22	Given an image and candidate generations, the figure shows the process for captioner fine-tuning by distilling from a CLIP-based evaluator. (Moratelli et al.).	42
3.23	Intermediate Label Conditioned Captioning: Architecture	43
3.24	Early Label Conditioned Captioning: Architecture	44
3.25	Late Label Conditioned Captioning: Architecture	45

3.26	Self-critical sequence training (SCST). The weight put on words of a sampled sentence from the model is determined by the difference between the reward for the sampled sentence and the reward obtained by the estimated sentence under the test-time inference procedure (greedy inference depicted). This harmonises learning with the inference procedure and lowers the variance of the gradients, improving the training procedure, (Rennie et al.).	46
3.27	Beam and Greedy Search Methods	47
4.1	X-CapFusion Research Workflow	50
4.2	Predictions for a ROCO test set example.	53
4.3	Predictions for a MIMIC NLE test set example.	53
4.4	Predictions for a ROCO test set example.	55
4.5	Predictions for a MIMIC NLE test set example.	55
4.6	Architecture	56
4.7	Training Parameters for Text Decoder and Image Encoder Selection	56
4.8	Image Encoder and Text Decoder Selection Experiments: Epoch Loss Overview	56
4.9	Predictions for a MIMIC-CXR NLE test set example.	58
4.10	GradCam: Analysis of the last layer of the image encoders. (a) Image with caption: "Linear scarring or atelectasis in the right mid lung field is similar." (b) Microsoft Rad-DINO image encoder: <code>layer model.encoder.layer.11.output</code> . (c) Microsoft BioMedCLIP image encoder: <code>layer visual.trunk.blocks.4</code> . (d) ViT-B/32 image encoder: <code>layer clip_model.visual.transformer.resblocks.11</code>	59
4.11	Dataset Selection Experiments: Epoch Loss Overview.	61
4.12	Dataset Selection Predictions for MIMIC-CXR dataset	63
4.13	First Model Adaptations Schema	64
4.14	Adaptations: Epoch Loss Overview	66
4.15	MIMI CXR Test Set Example	67
4.16	Second Model Adaptations Schema	69
4.17	Epoch Loss Overview - 10 epochs - CIDEr Reward RL Captioning	70
4.18	Epoch Loss Overview - 20 epochs - FocalLoss LabelPenalty	70
4.19	Epoch Loss Overview - 20 epochs	70
4.20	MIMI CXR Test Set Example	72
4.21	Full FineTunning Captioning Experiment: Gradient Analysis during training	73
4.22	Attention score (relative epoch 0)	74
4.23	Attention score (absolute)	74
4.24	Label Conditioned Captioning Models: Loss Overview	75
4.25	Decoded Soft Prompt examples: Early Label Conditioned Captioning	76
4.26	MIMIC CXR Test Set Example	77
4.27	NIH Test Set Example	78
D.1	Classifier Epoch Loss Overview	95

List of Tables

2.1	Summary of Fusion Methods	11
2.2	Table of a comparative analysis of the surveys on medical report generation.	15
2.3	Resume of the multimodal datasets for medical domain.	17
2.4	Table of a comparative analysis of image captioning research.	18
3.1	Comparison between BERT and GPT models.	29
3.2	Summary of the images split into training, validation, and test sets. [42].	32
3.3	Comparison of training methods and their architectures.	47
4.1	Baseline Models: Zero-shot Prediction Metrics in ROCO and MIMIC-CXR NLE datasets	51
4.2	Baseline Models: Few-shot Prediction Metrics in ROCO and MIMIC-CXR NLE datasets	54
4.3	Image Encoders and Text Decoders Pairs	55
4.4	Image Encoder and Text Decoder Selection Experiments on MIMIC-CXR NLE Dataset.	57
4.5	Dataset Selection Prediction Metrics in MIMIC-CXR Dataset	62
4.6	Model Adaptations Predictions metrics in MIMIC-CXR Regex Dataset	67
4.7	Model Adaptations Predictions metrics in MIMIC-CXR Regex Dataset	71
4.8	Ablation study: Predictions metrics in MIMIC-CXR Regex Dataset	76
4.9	Ablation Study: Predictions metrics in NIH Dataset	78
A.1	ClipCap Parameters	86
A.2	BLIP Parameters	86
A.3	MediCap Parameters	86
A.4	DiCO Parameters	86
A.5	ClipCap Parameters	87
A.6	MediCap Parameters	87
A.7	BLIP Parameters	87
A.8	DiCO Parameters	87
B.1	Model Training Parameters	89
C.1	Training Parameters for Dataset Selection	91
C.2	Training Parameters for Adaptations	93
C.3	Training Parameters for Adaptations	94
D.1	Classifier Evaluation Metrics in MIMIC-CXR NLE Dataset.	96
D.2	Training Parameters for Label Conditioned Models	97

Acronyms

AI Artificial Intelligence. [xvii](#), [1](#), [2](#), [3](#), [8](#), [14](#), [17](#)

BERT Bidirectional Encoder Representations from Transformers. [xviii](#), [xix](#), [3](#), [15](#), [19](#), [28](#), [29](#), [31](#), [35](#), [37](#), [38](#), [41](#), [50](#), [51](#), [57](#), [62](#), [75](#), [79](#)

BLEU BiLingual Evaluation Understudy. [35](#), [36](#), [37](#), [39](#), [50](#), [76](#), [79](#)

CIDEr Consensus-based Image Description Evaluation. [36](#), [38](#), [39](#), [50](#), [54](#), [63](#), [67](#), [69](#), [71](#), [72](#), [76](#), [79](#)

CLIP Contrastive Language-Image Pre-Training. [xviii](#), [3](#), [12](#), [14](#), [29](#), [30](#), [39](#)

CNN Convolutional Neural Network. [15](#), [18](#), [24](#), [25](#), [65](#), [66](#), [67](#), [68](#), [69](#), [70](#), [71](#), [73](#), [75](#), [82](#)

CT Computed Tomography. [1](#), [33](#)

DeepLIFT Deep Learning Important FeaTure. [21](#)

GPT Generative PreTrained Transformer. [4](#), [28](#), [29](#)

Grad-CAM Gradient-Weighted Class Activation Mapping. [21](#)

GRU Gated Recurrent Unit. [18](#)

LIME Local Interpretable Model-agnostic Explanations. [21](#)

LLM Large Language Models. [19](#), [22](#)

LM Language Modeling. [29](#)

LSTM Long-Short Term Memory. [18](#), [27](#), [28](#)

METEOR Metric for Evaluation of Translation with Explicit ORdering. 36, 37, 38, 39, 50, 79

MLM Masked Language Modeling. 29

MLP Multilayer Perceptron. 19, 40

MRI Magnetic Resonance Imaging. 1, 33

NLE Natural Language Explanation. xviii, xx, xxi, 31, 32, 39, 51, 52, 53, 54, 56, 57, 58, 60, 62, 63, 79, 81, 82, 92, 96

NLP Natural Language Preprocessing. xiv, 2, 26, 32, 36, 75

NSP Next Sentence Prediction. 29

PET Positron Emission Tomography. 33

RadCliQ Radiology Report Clinical Quality. 35, 36

RAG Retrieval-Augmented Generation. 21

RGB Red-Green-Blue. 15, 16

RL Reinforcement Learning. xvii, 20

RNN Recurrent Neural Network. 15, 18, 27, 28

ROCO Radiology Objects in COntext. xviii, 5, 30, 33, 34, 39, 51, 52, 54, 79, 81, 85

ROUGE Recall-Oriented Understudy for Gisting Evaluation. 36, 37, 39, 50, 54

SCST Self-Critical Sequence Training. 41, 42, 50, 51, 53

SOTA state-of-the-art. 15, 23, 49, 50, 63, 80, 81, 82

TF Teacher Forcing. 41

TF-IDF term frequency-inverse document frequency. 38

VIT Visual Transformers. 3, 4, 19, 26, 27

VQA Visual Question Answer. 20, 25, 40

Chapter 1

Introduction

This chapter introduces the context and motivation for this work, providing an overview of the problem description and objectives, and ending with a description of the document structure.

1.1 Introduction

In recent years, the intersection of Artificial Intelligence (AI) and medicine has shown great promise in transforming healthcare practices. A significant area of advancement is medical imaging, where AI-driven technologies have revolutionised diagnostics, treatment planning, and patient care. Among these technologies, multimodal AI stands out as a promising approach, integrating multiple sources of information such as images and text ([Zhang et al.](#)).

Image Captioning systems have emerged in recent years due to the need for automatic, accurate and interpretable image descriptions. While medical imaging techniques like Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, and X-rays, provide detailed anatomical and pathological information, interpreting these images can be complex and requires specialized knowledge ([Hussain et al.](#)). By automatically generating accurate and informative captions, multimodal AI systems can assist healthcare professionals in making faster and more informed decisions (Figure 1.1).

Moreover, the context of rising healthcare demands and the increasing volume of medical data further underscores the necessity for advanced AI solutions. Traditional image analysis and reporting methods can be time-consuming and prone to variability ([Madabhushi and Lee](#)). Multimodal AI enhances image interpretation efficiency and potentially

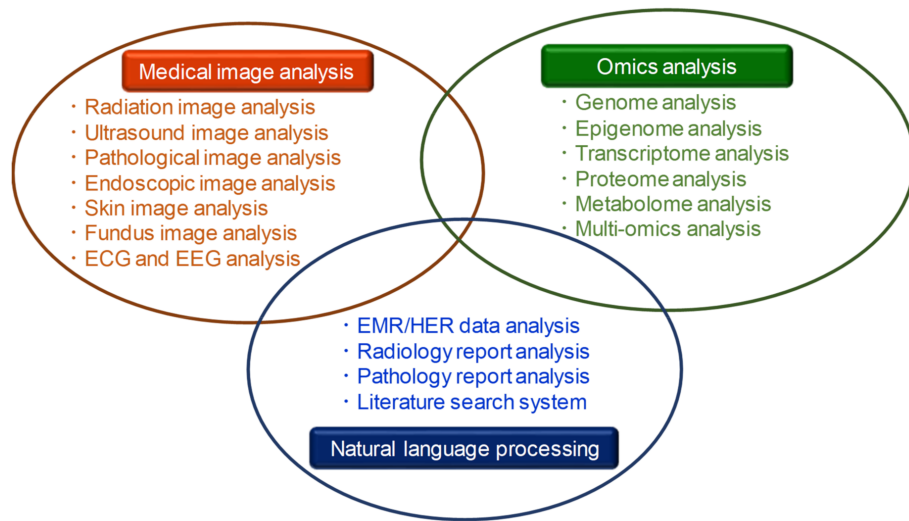


FIGURE 1.1: Diagram with three main areas in which AI technology in medical field is currently being implemented: medical image analysis, omics analysis, and natural language processing (Hamamoto).

reduces diagnostic errors by providing standardized, consistent descriptions based on visual and contextual information (Rao et al.).

Furthermore, the evolution of deep learning techniques, particularly in Natural Language Preprocessing (NLP) Vaswani, and computer vision Dosovitskiy et al., has paved the way for multimodal AI applications Wang et al.. These advancements enable systems to recognize image patterns and, understand and generate human-like descriptions. This capability holds immense potential for enhancing medical imaging workflows, automatic generation of reports Li et al., enabling remote and accurate diagnostics, improving communication between healthcare providers and patients, and ultimately improving patient outcomes.

1.2 Motivation and Contextualization

Medical image captioning aims to improve healthcare outcomes and address key challenges in clinical practice, such as enhancing clinical decision-making by assisting or providing preliminary descriptions, helping identify critical findings and reducing diagnostic errors (Reale-Nosei et al.).

Captioning systems, whose general pipeline is depicted in Figure 1.2, can help reduce variability in reporting styles among radiologists, ensuring standardized, clear, and concise descriptions that improve communication and understanding across clinical teams,

facilitating data analysis by allowing researchers to extract and summarize information from large medical image datasets, enabling efficient analysis for studies in epidemiology, pathology, and treatment outcomes, and make medical imaging more accessible to non-specialists by providing clear explanations that improve understanding of medical conditions.

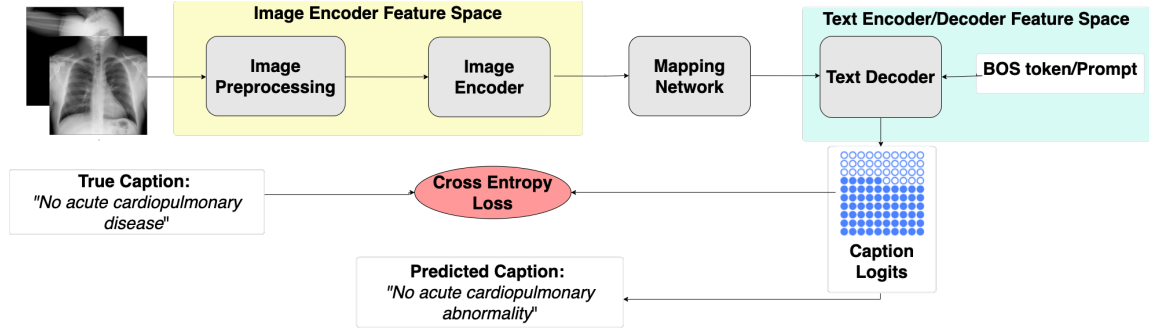


FIGURE 1.2: Diagram with the general Generative Based Image Caption Generation Pipeline with Mapping.

Although medical image captioning can improve many aspects of medical practice, and despite its potential, it still faces challenges such as domain-specific language requirements, dataset limitations, and ensuring clinical accuracy. These challenges drive innovation in using advanced methods like transformers, multimodal fusion techniques, and domain-specific regularization (He et al.), ensuring this field continues to evolve and positively impact healthcare systems. Multimodal AI systems are more resilient to noise and missing data, they can rely on other modalities to maintain performance if one modality is unreliable or unavailable (Kim and Chung), and it allows performing a new set of tasks such as visual-question answering, image captioning, multimodal retrieval and recommendation systems.

1.3 Methodology Employed

The methodology employs state-of-the-art architectures, such as Transformers, Bidirectional Encoder Representations from Transformers (BERT), Visual Transformers (ViT), and Contrastive Language-Image Pre-Training (CLIP) models, to develop a medical image captioning system. It begins with a literature review to survey existing approaches, identify strengths and limitations, and highlight challenges like dataset constraints, domain-specific language complexity, and benchmarking gaps.

The system is designed with modular components, using transformer-based encoders (e.g., ViT, Swin Transformers) and transformer-based text decoder (e.g., Generative Pre-Trained Transformer (GPT)). Those are connected by transformer-based mapping networks to align each modality embeddings and multi-label classes incorporated to enhance the performance. Regularization techniques, including adversarial training and soft prompt class-oriented cues, are applied to improve robustness and clinical relevance.

Implementation involves dataset preparation, testing various encoder-decoder configurations, and benchmarking against existing baselines. An ablation study evaluates the contribution of different components and inclusion of labels through multiple fusion techniques, while generalization is assessed across multiple datasets. This approach delivers a robust, flexible system addressing current challenges and advancing the field.

1.4 Objectives

This work aims to design and implement a solution to address the challenges posed by the increasing need for effective medical imaging caption systems. Our objective is twofold: first, to explore and analyze the current literature in this field, and second, to develop a medical image captioning model that generates accurate and interpretable descriptions.

To achieve this, we propose a method based on pre-trained unimodal backbones, focusing on training only the projection layers to align image and text features in a shared latent space. Furthermore, label-conditioned captioning is tested across multiple traditional and domain specific metrics. This approach is intended to enhance caption quality and streamline training processes for specific datasets. By studying and improving the alignment of image and text features, we aim to create a robust, lightweight multimodal system capable of efficiently delivering precise and meaningful medical image captions. Our contributions are summarized as:

- A tailored literature review and taxonomy of state-of-the-art multimodal architectures applied to medical image captioning.
- A lightweight captioning model using pre-trained unimodal backbones with frozen weights, training only the multimodal fusion layers.
- A projection-based alignment strategy for combining image and text features into a shared latent space.

- Experimental implementation of label-conditioned medical captioning models in three fusion topologies.
- Experimental evaluation on medical captioning benchmarks (Radiology Objects in COntext (ROCO), MIMIC-CXR) with competitive performance.
- A detailed ablation study on the impact of labels inclusion and alignment strategies.
- A contribution to interpretability by generating captions that align with medically relevant terminology.
- Public release of the code and models to foster reproducibility and future work in the community.

1.5 Outline of the Thesis

This thesis is structured as follows:

- Chapter 2 concisely presents a literature review regarding Image Caption, namely, focused on the medical domain and multimodal systems.
- Chapter 3 presents a detailed description of the methodology to be employed, complemented with the mathematical formulation of models and main components, aiming to identify the most prominent set of models and possible changes to be applied.
- Chapter 4 presents the implementation and more meticulous details of the methodology to be employed, as well as a detailed analysis of the conducted experiences and a discussion of the results.
- Chapter 5 discusses the overall findings and sets the conclusions and future work.

Chapter 2

Literature Review

This chapter presents a comprehensive review of the current research in medical image captioning, an interdisciplinary field combining computer vision and natural language processing to generate clinically relevant textual descriptions of medical images.

2.1 Multimodal

Multimodal refers to systems that integrate and process multiple forms of data, such as text, images, or audio, to achieve a more comprehensive understanding and generate more robust outputs than models trained on a single data type. These models leverage the strengths of different modalities to enhance knowledge and provide more prosperous and accurate outputs. For example, Google’s multimodal system, Gemini ([Team](#)), can receive a photo of a plate of cookies and generate a written recipe as a response and vice versa.

2.1.1 Multimodal Characteristics

Multimodal can be described in three fundamental characteristics: heterogeneity, connections, and interactions ([Liang et al.](#)).

- **Heterogeneity** refers to the diverse qualities, structures and representations of modalities. A text description of an event will fundamentally differ in quality, structure and representation from a photograph of the same event.

- **Connections** refers to the complementary information shared between different modalities. These connections may be reflected in statistical similarities or semantic correspondence.
- **Interactions** refers how modality elements interact to give rise to new information when integrated together for task inference.

The core engineering challenge for multimodal lies in effectively integrating and processing diverse types of data to create models that can leverage the strengths of each modality while overcoming their limitations. The work of [Liang et al.](#) also presents several challenges: representation, alignment, reasoning, generation, transference and quantification (Figure 2.1).

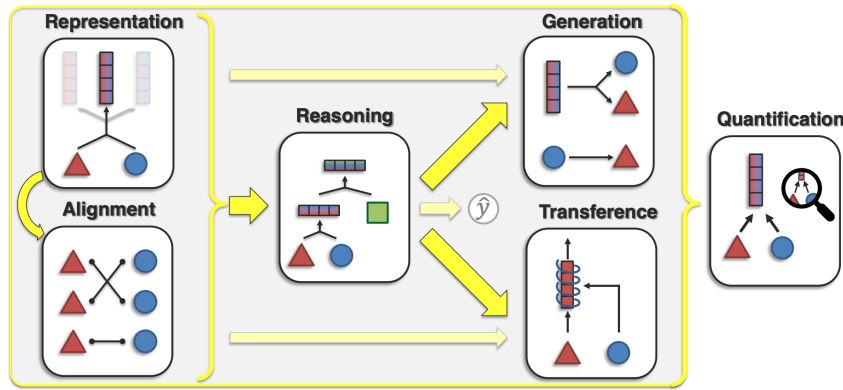


FIGURE 2.1: Core research challenges in multimodal learning ([Liang et al., 2023](#)).

Representation studies how to represent and summarize multimodal data to reflect the heterogeneity and interconnections between individual modality elements. **Alignment** aims to identify the connections and interactions across all aspects, and **Reasoning** aims to compose knowledge from multimodal evidence, usually through multiple inferential steps for a task. **Generation**, on the other hand, involves learning a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence, while **Transference** aims to transfer knowledge between modalities and their representations. **Quantification** involves empirical and theoretical studies to understand the multimodal learning process better.

2.1.1.1 Multimodal Fusion Approaches

In multimodal Artificial Intelligence (AI), fusion techniques are crucial for integrating information from different modalities (text, images, audio, etc.) to enhance understanding

and decision-making. It integrates data from various sources, such as text, images, audio, video, and sensor readings, to better represent the underlying information (Figure 2.2).

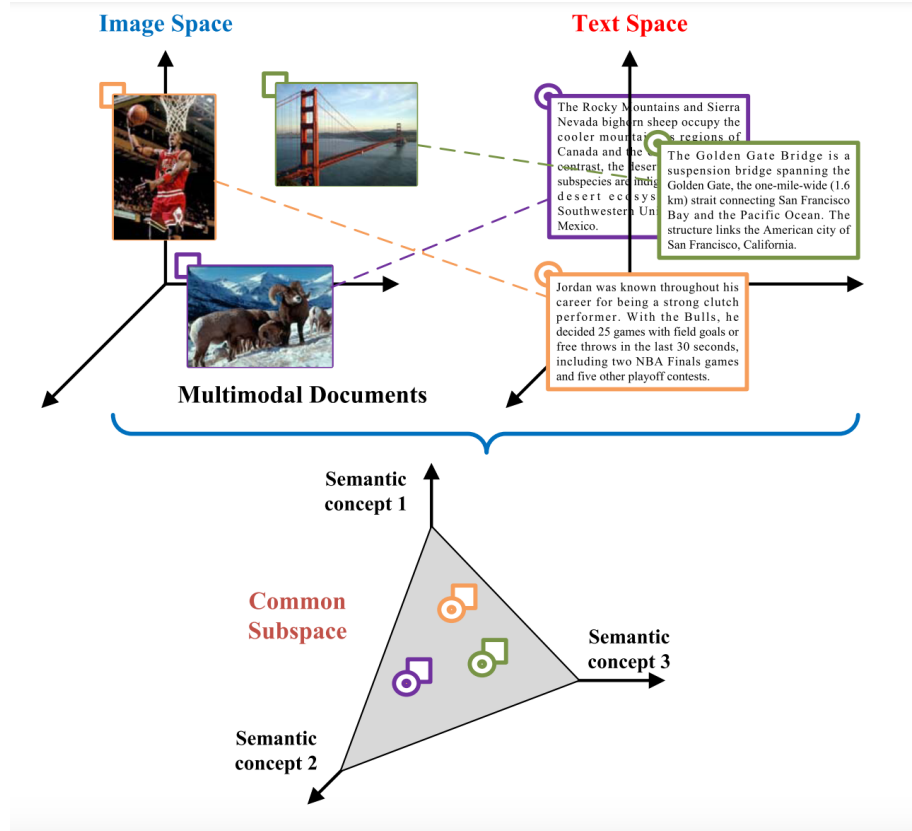


FIGURE 2.2: Schematic of the common subspace learning, which aims to project the heterogeneous data of different modalities into a common subspace, where the multimodal data with similar semantics will be represented by a similar vector (Guo et al.).

The fusion techniques can be categorized following the taxonomy employed in Guo et al., and Boulahia et al., depicted in Figure 2.3:

- **Early Fusion:** In this approach, raw data from different modalities is combined at the input level before being fed into a model. The advantage of this strategy is that we do not have to perform dedicated processing for each modality. We fuse them all at the input level and process the fused representation through the model. The downside is that raw input data may not contain rich semantic information, such that the model cannot capture complex interactions between the modalities, limiting the benefit the model can derive from the multimodal fusion.
- **Intermediate fusion:** This approach combines data from different modalities at various intermediate processing stages within a model architecture. It is the most widely used fusion strategy and involves processing each modality into a latent

representation, fusing them, and then doing some final processing to produce the output scores. Fusing the latent representations is an implementation detail, but some options include concatenation, element-wise addition, or even attention mechanisms. The advantage of intermediate fusion is that the model can learn rich interactions between each modality because they are first converted into a machine-understandable representation. The downside to this approach is that individual processing is required for each modality before they can be fused. This may not be a huge issue because of the availability of pre-trained embeddings for images, text and audio. Still, it adds extra processing time and can impact inference speed.

- **Late fusion** is closer to an ensemble model. Each modality is processed independently with its models, and their outputs are combined later. The advantage of late fusion is its simplicity and isolation. Each model gets to learn super-rich information on its modality. The downside is that the system cannot learn complex modal interactions and thus does not benefit directly from the complementary information each modality might offer.

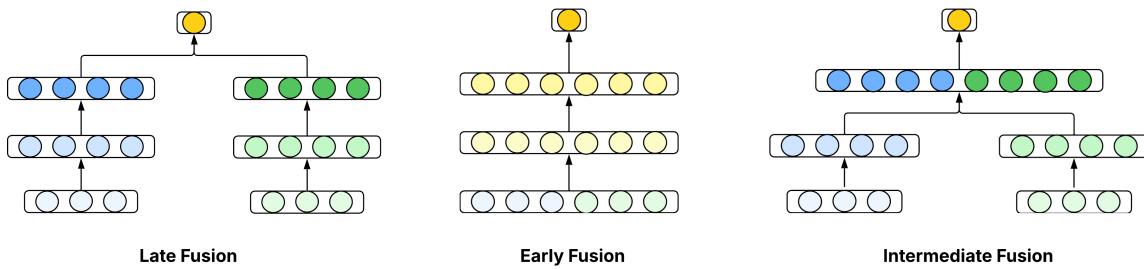


FIGURE 2.3: Multimodal Fusion Techniques.

Hybrid fusion techniques that combine early, intermediate, and late elements, described in Table 2.1, are also explored to optimize multimodal integration based on specific use cases.

TABLE 2.1: Summary of Fusion Methods

Fusion Method	Advantages	Disadvantages
Early Fusion Barnum et al. Team Barnum et al. Sun et al. Miranda-Belmonte and Lopez-Monroy	Does not need dedicated processing for each modality.	Raw input data may not contain rich semantic information. Can be computationally intensive and prone to overfitting with high-dimensional data.
Intermediate Fusion Li et al. Mokady et al. Nicolson et al. Wang et al.	The model can learn rich interactions between each modality.	Individual processing is required, and there is difficulty in synchronising different feature spaces.
Late Fusion Radford et al. Li et al. Bird et al.	Simple to perform and isolation of the outcomes of each modality.	May lose modality interactions and require additional processing to align outputs.

Each approach has its strengths and trade-offs, and the choice of fusion strategy should align with the application’s specific requirements, such as model efficiency, data complexity, and the depth of multimodal understanding needed.

2.1.2 Multimodal Image Captioning Approaches

The main categories of existing image captioning methods include template-based image captioning, retrieval-based image captioning, and novel caption generation, depicted in Figure 2.4.

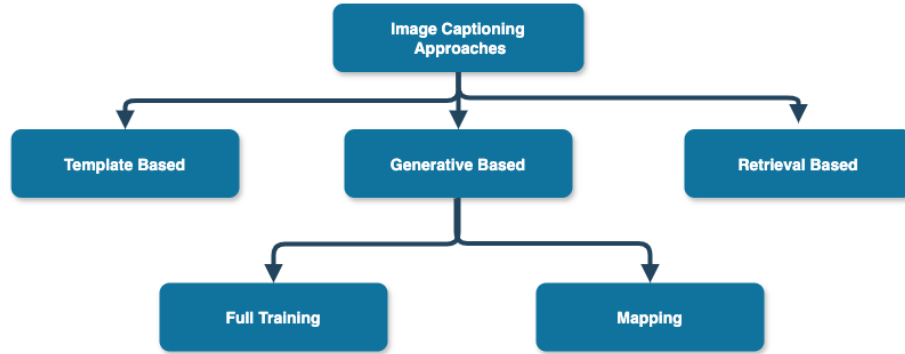


FIGURE 2.4: Image Captioning Methods.

- **Template-based approaches** have fixed templates with several blank slots to generate captions. In these approaches, different objects, attributes, and actions are detected first, and then the blank spaces in the templates are filled. For example, Farhadi et al. uses a triplet of scene elements to fill the template slots for generating image captions. Template-based methods can generate grammatically correct captions. However, templates are predefined and cannot generate variable-length captions.

- **In retrieval-based approaches**, captions are retrieved from a set of existing captions. First they find visually similar images with captions from the training dataset. These captions are called candidate captions. The captions for the query image are selected from the captions pool. Contrastive Language-Image Pre-Training (CLIP) [Radford et al.](#), can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing the task, being used as a retrieval-based approach for image captioning. These methods produce general and syntactically correct captions. However, they require a set of textual captions to choose from and do not allow specific and free captions to be generated.
- **Captions Generation** is an approach where captions are generated from a visual or multimodal space. A typical framework consists of an image encoder to analyze the visual content of the image and a language decoder to create captions for that image. These methods can generate new captions for each semantically more accurate picture than previous approaches and have more degrees of freedom. Therefore, deep learning-based novel image caption-generating methods are our primary focus in this work.

There are two main research branches of the caption generation approach as depicted in Figure 2.5:

- **Full Training**: Use brand new image encoder architecture and text decoder to train all the models from scratch as employed by [Nicolson et al.](#). This is a heavy model that needs more computational power.
- **Mapping**: Use a pre-trained image encoder and text decoder, and use a trainable mapping network to join them. This lightweight approach was adopted in the work of [Mokady et al.](#), where they use a transformer-based mapping network to collapse image features into the text decoder embedding space as well as possible. Then, the model's performance depends on the ability of the image encoder to extract relevant features from the image and the ability of the text decoder to generate consistent and organized captions from the embeddings.

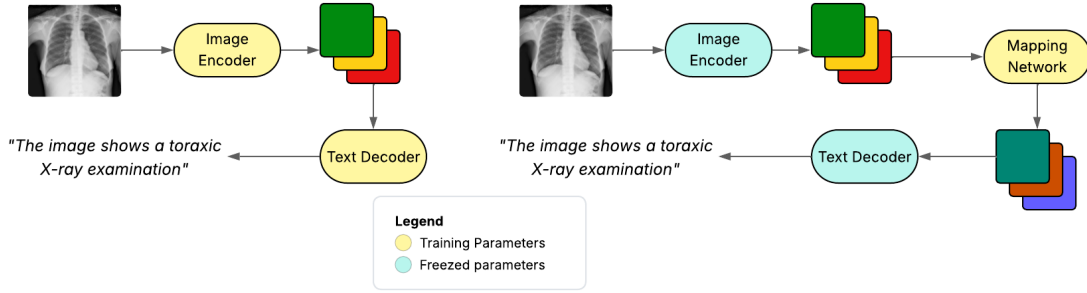


FIGURE 2.5: Generative Approach: Full Training (left) and Mapping (right).

2.2 Literature Search

During the last few years, computer vision and natural language fusion have created the captivating field of image captioning that involves understanding the semantics of the image, which requires understanding the main objects, their various attributes, poses, and their interactions in an image. It also needs to infer the underlying semantic meanings to generate meaningful captions.

2.2.1 Search Methodology

To perform a focused literature review, we employed a top-down approach (Figure 2.6) in multiple leading research databases, including ArXiv, Litmaps and IEEE Xplore. The search aimed to identify seminal works and recent advancements in the field of medical image captioning.

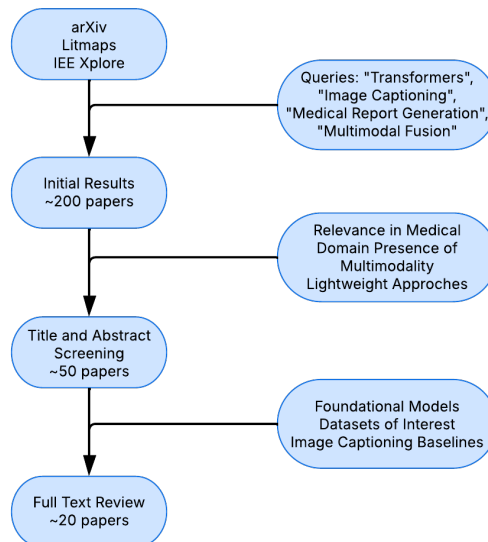


FIGURE 2.6: Search Methodology: Top-down

We constructed targeted search queries using keywords such as ‘Transformers’, ‘Image Captioning’, ‘Medical Report Generation’, and ‘Multimodal Fusion’. These queries were designed to capture foundational works and recent innovations within the domain.

Two pivotal works, Transformer [Vaswani](#) and CLIP [Radford et al.](#), were selected as benchmarks due to their foundational role in modern multimodal learning. Using Litmaps, we identified related works (Figure 2.7) applied to medical image captioning, which led us to discover several significant contributions, including ClipCap by [Mokady et al.](#). This process allowed to systematically map the relationship between foundational technologies and their adaptation for specific applications in the medical field.

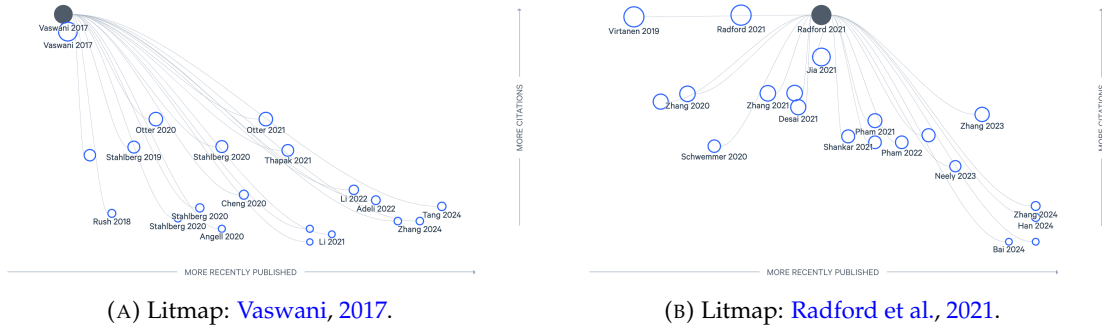


FIGURE 2.7: Litmap: Foundational Models.

The first graph revolves around “Attention Is All You Need” by [Vaswani](#), the seminal work introducing the Transformer architecture. This groundbreaking model became the foundation of modern AI, influencing diverse fields, from natural language processing to computer vision. The map showcases an extensive lineage of derivative research, with key developments like [Otter et al.](#) and [Li et al.](#) contributing to incremental advancements or specialized adaptations of the Transformer.

The second graph centred on the transformative work by [Radford et al.](#), introduces CLIP, a foundational advancement in multimodal contrastive learning. CLIP’s core innovation lies in its ability to align visual and textual representations through large-scale pre-training on diverse internet data, marking a paradigm shift in multimodal AI by demonstrating how contrastive learning could bridge modalities effectively. It has inspired a cascade of research, linking foundational works like [Zhang et al.](#), which explored applications in medical image-text pairs, and contemporaneous studies such as [Zhao et al.](#), that apply the technology to diffusion models.

2.2.2 Review of Existing Surveys

Several recent studies concerning surveys for multimodal data fusion can be found in the literature (Table 2.2).

[Zhao et al., 2024](#) provides a focused study of multimodal data fusion based on encoder-decoder approach methods, Attention mechanism, Graph-based neural networks, and Generative Neural Networks, among others. However, it focuses on vision-text and vision-sensors, which are standard applications, avoiding specific medical applications. [Gao et al.](#) presents a state-of-the-art (SOTA) review on multimodal neural networks, only focusing on object recognition in Red-Green-Blue (RGB) domains, mainly using Convolutional Neural Network (CNN). [Zhang et al.](#) presents a multimodal data fusion survey but concentrates only on the image segmentation tasks while [Abdu et al.](#) and [Zhu et al.](#) present a multimodal survey for sentiment analysis, with [Abdu et al.](#) focusing on deep learning approaches in combination with early, intermediate and late fusion.

TABLE 2.2: Table of a comparative analysis of the surveys on medical report generation.

Papers	Year	Multimodal	Modalities	Datasets	Benchmarking	Metrics
A survey on biomedical image captioning [77]	2019	✓	Radiology + Text. (5 works)	✓	✓	✓
Automatic medical image interpretation: State of the art and future directions. [2]	2021	✓	Radiology + Text. (18 works)	✓	✓	✓
Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images [67]	2022	✓	Radiology, Histopathology, Occular + Text. (25 works)	✓	✓	✓
Diagnostic captioning: a survey. [78]	2022	✓	Radiology + Text. (19 works)	✓	✓	✓
A systematic review of deep learning-based research on radiology report generation. [62]	2023	✓	Radiology + Text. (22 works)	✓	✓	✓
Automatic captioning for medical imaging (MIC): a rapid literature review. [6]	2023	✓	Radiology + Text. (42 works)	✓	✓	✓
Deep image captioning: A review of methods, trends and future challenges. [125]	2023	✓	General Images + Text.	✓	✓	✓
A survey on automatic generation of medical imaging reports based on deep learning [76]	2023	✓	Radiology + Text. (31 works)	✓	✓	✓
From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. [90]	2024	✓	Radiology + Text. (15 works)	✓	✓	

The works of [Bayoudh et al.](#), [Gao et al.](#), provide a survey focused on multimodal data fusion that includes encoder-decoder architectures, CNN methods, Graph-Neural Network methods and Recurrent Neural Network (RNN) ones, but novel transformer and attention mechanisms, present in the large language models such as Bidirectional Encoder

Representations from Transformers (BERT), are not included. A medical domain survey such as [Shamshad et al.](#) focuses only on transformer architectures for several tasks, such as classification, segmentation, registration and restoration. However, it fails to identify relevant multimodal work approaches that have established benchmarks for multimodal medical domains. Other surveys in multimodal, such as [Baltrušaitis et al.](#), [Gu et al.](#), [Hermessi et al.](#), [Li et al.](#) address multimodality, but most of them focus on combining different images from different colour spectra, such as RGB and X-Ray.

Despite these advancements, there remains room for improvement in accurately capturing the complexities of medical imaging and reporting, highlighting the need for more sophisticated multimodal approaches to enhance the field of automated medical image captioning. Therefore, the scope of this survey is to focus only on multimodal methods for medical image captioning. However, some pre-trained models for other downstream tasks, such as visual question answering, segmentation, and image/text retrieval, can also be updated for this task.

2.2.3 Multimodal Datasets

The most common datasets to train and assess the performance of medical caption generation are the MIMI-CXR ([Johnson et al.](#)), IU X-Ray (Indiana University Chest X-Ray Collection) ([Pavlopoulos et al.](#)), ROCO (Radiology Objects in COntext) ([Pelka et al.](#)), ImageCLEF Medical Caption Dataset, PEIR ([Pavlopoulos et al.](#)), MedPix ([Siragusa et al.](#)), Quilt-1M ([Ikezogwo et al.](#)), PadChest ([Bustos et al.](#)), MedICaT ([Subramanian et al.](#)), MedTrinity-25M ([Xie et al.](#)) described in Table 2.3. Most of the datasets above contain images on different modalities and their corresponding textual reports, enabling the performance of several multimodal downstream tasks such as visual question answers, image-to-text retrieval, image captioning, and report generation.

TABLE 2.3: Resume of the multimodal datasets for medical domain.

Dataset	Examples	Modality	Description
MIMIC-CXR [43]	372,000 images linked into almost 228 thousand reports written in English	X-ray.	Images of Chest-ray paired with text reports in English
PadChest [10]	160,868 chest x-rays, six different perspectives, coupled with 109,931 reports	X-ray.	Images of Chest-ray paired with text reports pairs in Spanish. The reports were labeled with 174 different radiographic findings, 19 differential diagnoses and 104 anatomic locations organized as a hierarchical taxonomy.
PEIR [77]	7,443 images along with their corresponding captions.	X-rays, CT scans, MRI, ultrasounds, histopathology.	Single descriptive sentence. Suitable for classification or organ identification. Caption describing the medical context, anatomical region, and specific pathology or condition.
ChexPert [40]	224,316 chest radiographs of 65,240 patients and associated radiologist interpretations for various pathologies.	X-ray.	Chest images, both frontal and lateral views, with interpretations for various pathology, including pneumonia, oedema, and atelectasis. Expansion from Chest-ray8 with more lesions added. Allows caption detection and generation.
ImageCLEF Caption dataset (2017, 2018, 2019) [39].	Final version with 232,305 image-caption pairs, with 222,305 images allocated for training and 10,000 for testing.	X-rays, CT scans, MRI, ultrasounds.	Rich set of image-text pairs dataset, on various medical domains. Wide range of medical domains, simple to complex captions, from concept identification to caption or answer generation.
ROCO [80].	81,000 radiology images encompassing various medical imaging modalities, the dataset served as a crucial resource for the ImageCLEF 2015 Medical and other out-of-class images.	X-rays, CT scans, MRI, ultrasounds, histopathology.	Rich set of image-text pairs dataset, on various medical domains. A more recent version ROCov2 [94] includes a more extensive and curated collection of images and captions.
Quilt-1M [38] .	419,780 images followed with 768,826 text reports	Histopathology.	Complemented with focused ROI entities extracted from the associated text.
MedPix [102] .	9,000 topics, with 59,000 images and reports.	X-rays, CT scans, MRI, ultrasounds, histopathology, dermatology, and ophthalmology.	Covers different anatomical regions, diseases, and medical conditions, complemented with educational materials.
IU Chest X-Ray [77] .	7,470 chest X-ray images with 3,955 text reports.	X-rays.	Images are high quality, thoracic diseases and condition well documented.
MedICaT [103] .	217,060 images and 7507 sub captions from articles.	X-rays, CT scans, MRIs, ultrasounds, and microscopic images.	Wide range of anomalies mined from PubMed published articles.
PMedTrinity-25M [123].	32,799 open-ended questions from 4,998 pathology images.	X-rays, CT scans, MRIs, ultrasounds, PET, and microscopic images	Very Large set of modalities, and rich information.

Figure 2.8 displays the increased usage trend of the MIMIC-CXR dataset from 2020 to 2025, reinforcing the high demand for multimodal AI solutions in the medical domain.

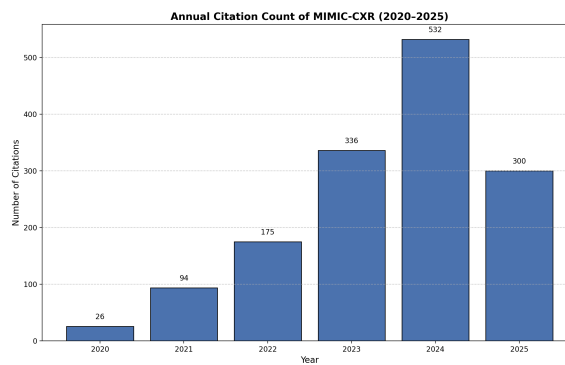


FIGURE 2.8: Trend of the dataset usage

2.2.4 Image Captioning Works

Solutions to accurately and efficiently generate medical text captions have been widely explored. Thus, numerous surveys listed in Table 2.2 identify and evaluate most existing

works; however, except for the most recent survey [Reale-Nosei et al.](#), it lacks the complete multimodal works that are now the established benchmark in image caption generation (Table 2.4).

TABLE 2.4: Table of a comparative analysis of image captioning research.

Papers	Year	Datasets	Image Encoder	Text Decoder	Metrics
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [124]	2016	Flickr8k, Flickr30k and MS COCO	CNN	LSTM	METEOR and BLUE
Unified Vision-Language Pre-Training for Image Captioning and VQA [136]	2019	COCO Flickr30k and VQA 2.0	Captions, Captions, Faster R-CNN	BERT	BLUE, METEOR, CIDEr and SPICE
ClipCap: CLIP Prefix for Image Captioning: [69]	2021	COCO-captions, no-caps, Conceptual Captions	CLIP	GPT2	BLEU, METEOR, CIDEr, SPICE, ROUGE-L
Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning [32]	2021	IU-Xrays, MIMIC-CXR	VGG-16, ResNet-152	LSTM	BLUE, METEOR, CIDEr and ROUGE-L
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation [56]	2022	NoCaps, COCO	ViT	BERT	SPICE, BLUE and CIDEr
LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day [51]	2023	VQA-RAD, SLAKE, PathVQA	CLIP, BioMedClip	LLaVA, Vicuna	Accuracy(closed-set questions) and Recall (open-set questions)
Medical image captioning via generative pre-trained-language model [95]	2023	MIMIC-CXR, Open-i IU Xray, MS-COCO	SAT	SAT, GPT3	BLEU, CIDEr, ROUGE-L
Sam-Guided Enhanced Fine-Grained Encoding with Mixed Semantic Learning for Medical Image Captioning [133]	2023	COCO, ROCO, MedI-CaT	ViT Encoder	OPT LLM	BLEU, METEOR, ROUGE L, CIDEr, BERTSCORE, BARTSCORE, and BLEURT
Longitudinal Data and a Semantic Similarity Reward for Chest X-ray Report Generation [73]	2024	MIMIC-CXR, Open-i IU Xray	CvT-21	BERT	CheXbert, RadGraph ER, BLEU, BERTScore F1, F14, CIDEr, ROUGE-L, CXR-BERT
VILA: On Pre-training for Visual Language Models [61]	2024	OKVQA, TextVQA, COCO, Flickr	ViT	Llama-2	accuracy and CIDEr
Uterine Ultrasound Image Captioning Using Deep Learning Techniques [9]	2024	Zenodo, Uterine fibroid ultrasound images	InceptionV3, DenseNet201.	BiGRU	BLUE and ROUGE-L
RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models [122]	2024	MIMIC-CXR, IUxray, HarvardFairVLMed	LLaVA-MED-1.5 7B, ResNet-50	LLaVA-MED-1.5 7B, bio-BioClinicalBERT	BLUE, ROUGE-L and METEOR
MMED-RAG: Versatile Multimodal RAG System for Medical Vision Language Models [121]	2024	IU-Xray, MIMIC-CXR, Havard-FairVLMed, QUILT-1M, PMC-OA	LLaVA-MED-1.5 with MMED-RAG	LLaVA-MED-1.5 with MMED-RAG	BLUE, ROUGE-L and METEOR
CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models [120]	2024	MIMIC-CXR, IU-Xray, HarvardFairVLMed, PMC-OA, HAM10000, OL3I, OmniMedVQA.	LLaVA-Med, Med-Flamingo, MedVInT RadFM, LLaVA-v1.6, Qwen-VL-Chat	LLaVA-Med, Med-Flamingo, MedVInT RadFM, LLaVA-v1.6, Qwen-VL-Chat	Accuracy and over-confident ratio

Initially, most methodologies developed for medical image caption generation or question answering employed RNNs as text decoders. Several surveys identify works that use traditional CNNs such as VGGNet [Simonyan and Zisserman](#) (VGG16 and VGG19), ResNet [He et al.](#) (ResNet50, ResNet101, and ResNet152), Inception Network [Szegedy et al.](#) (InceptionV3), DenseNet [Huang et al.](#) (DenseNet121, DenseNet169, and DenseNet201), Efficient Net [Tan and Le](#) (EfficientNetB0, EfficientNetB1) and MobileNet [Howard et al.](#) as image encoders, paired with Long-Short Term Memory (LSTM) [116] or Gated Recurrent Unit (GRU) [99] networks as text decoders for caption generation like [Messina](#)

et al., Pavlopoulos et al..

While the advent of deep learning has brought significant improvements to medical report generation, the main conclusions suggest that inherent model constraints and the lack of true multimodal training still limit the performance of evaluated models.

With the appearance of attention mechanisms in 2017 by Vaswani et al., image captioning models took another path using transformer-based architectures. Examples of transformer based image encoders are CLIP Radford et al., Visual Transformers (ViT) Dosovitskiy et al., MedViT Manzari et al., Swin Transformer Liu et al., UniFormer Li et al.. Large Language Models transformer based can be used as text decoders like GPT Radford and Narasimhan, BART Lewis et al., XLNet Yang et al., MedGPT Kraljevic et al., BERT Devlin et al., T5 Raffel et al..

Mokady et al. proposed ClipCap, and it is tested a transformer-based and an Multi-layer Perceptron (MLP) mapping network to project image embedding space into Large Language Models (LLM)s embedding space. According to Lin et al., freezing LLMs during pre-training can achieve decent zero-shot performance but lack in-context learning capability, which requires unfreezing it.

Li et al. proposed BLIP to effectively utilize the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. A strong generalization ability was demonstrated when directly transferred to video language tasks in a zero-shot approach.

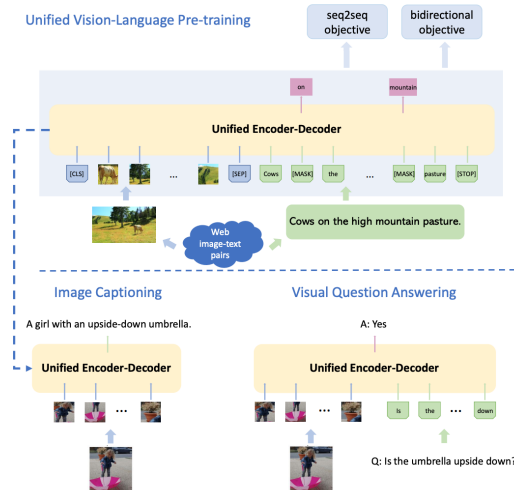


FIGURE 2.9: A unified encoder-decoder model for general vision-language pre-training. The pre-trained model is then fine-tuned for image captioning and visual question answering (Zhou et al.)

[Zhou et al.](#) proposes Unified Vision-Language Pre-Training for Image Captioning and Visual Question Answer (VQA) that can be finetuned for either vision-language generation (e.g., image captioning) or understanding (e.g., visual question answering) tasks, and uses a shared multi-layer transformer network for both encoding and decoding, which differs from many existing methods where the encoder and decoder are implemented using separate models, achieving competitive results. The model, described in Figure 2.9, is pre-trained on large amounts of image-caption pairs and optimised for two unsupervised vision-language prediction tasks: directionality and sequence-to-sequence (Seq2Seq) masked language prediction.

[Hou et al.](#) propose a framework involving three components: encoder, decoder, and reward module. While previous medical-report-generation models are trained mainly by minimizing the cross-entropy loss or further optimizing the standard image-captioning metrics, ignoring diagnostic accuracy, which should be the first consideration in this area, this paper, inspired by the generative adversarial network, proposes an adversarial reinforcement learning approach for report generation of chest x-ray images considering diagnostic accuracy and language fluency as depicted in Figure 2.10.

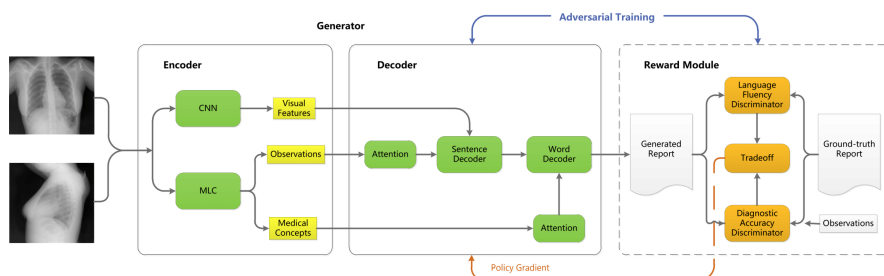


FIGURE 2.10: Overview of proposed framework involving three components: encoder, decoder, and reward module. The encoder comprises two branches that extract visual and semantic features for the decoder separately. The decoder generates a report hierarchically. Note that the decoding process is not unfolded due to space limitations. In adversarial training, the generator is trained by Reinforcement Learning (RL) using rewards offered by the discriminators. Simultaneously, the discriminators are updated by the maximum-likelihood estimation approach ([Hou et al.](#)).

[Zhang et al.](#) presents a novel medical image captioning method guided by the Segment Anything Model (SAM) to enable enhanced encoding with both general and detailed feature extraction. It comprises a dual-encoder architecture that enables capturing information with different granularities in medical images. A distinctive pre-training strategy complements both encoders, showcasing excellent synergy while preserving the overall and detailed semantic knowledge from the two encoders.

Li et al. proposes a cost-efficient approach for training a vision-language conversational assistant that can answer open-ended research questions of biomedical images. The main contribution is to leverage a large-scale, broad-coverage biomedical figure-caption dataset extracted from PubMed Central and use GPT-4 to self-instruct open-ended instruction-following data from the captions, and then fine-tune a large general-domain vision-language model using a novel curriculum learning method.

Xia et al. presents a multimodal Retrieval-Augmented Generation (RAG) system, MMed-RAG, designed to enhance the factuality of Med-LVLMs (Figure 2.11). MMed-RAG can achieve an average improvement of 43.8% in the factual accuracy of Med-LVLMs. This study concludes that a large medical visual language model tends to rely more heavily on the text while ignoring visual information. When retrieval information is introduced, the original Med-LVLM focuses more on the retrieved answers, even if the content is incorrect.

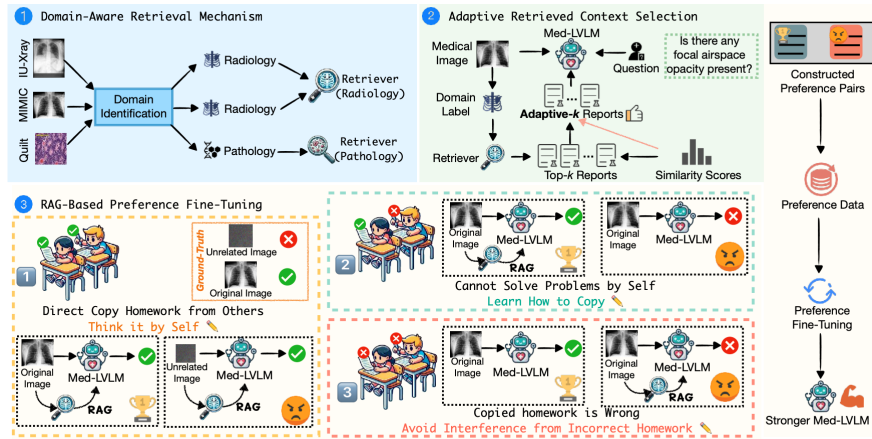


FIGURE 2.11: Overview of MMed-RAG, a versatile factual multimodal RAG system designed to enhance the reliability of Med-LVLMs. It introduces a domain-aware retrieval mechanism that effectively handles different domains of medical images by selecting suitable retrieval models. Additionally, it uses an adaptive context selection approach to determine the optimal number of retrieved contexts. It employs preference fine-tuning to improve both cross-modality and overall alignment (Xia et al.).

Explainable Artificial Intelligence (XAI) is a perspective explored by Messina et al., with many works relying on standard techniques such as Gradient-Weighted Class Activation Mapping (Grad-CAM) Selvaraju et al., Guided Backpropagation Nie et al. or Deep Learning Important Feature (DeepLIFT) (Li et al.), and even Local Interpretable Model-agnostic Explanations (LIME) (Garreau and Luxburg).

More novel-identified works explore the automatic generation of a natural language report to explain a classification task or present a counterfactual local explanation and

the text form. However, the identified approaches need to include an explanation of the generated report.

Dataset class imbalance is also a concern in multiple studies, as highlighted in the survey by [Messina et al.](#), posing a significant challenge in medical image captioning. Publicly accessible datasets in this field consistently exhibit a disproportionately low representation of abnormal cases [Tajbakhsh et al.](#). This issue is further exacerbated when considering individual diseases, where the sample sizes become even more limited. Such imbalances pose a substantial obstacle to model generalization, potentially compromising the ability of machine learning algorithms to accurately interpret and describe a wide range of medical conditions, particularly less common ones. This underrepresentation of abnormal cases in training data can lead to biased models that perform inadequately on crucial, albeit infrequent, medical scenarios.

The emergence of LLMs has brought attention to the hallucinations discussed in [Reale-Nosei et al.](#) when using the MIMIC-CXR medical caption dataset. It is a phenomenon where models generate plausible but factually incorrect information. To tackle those problems, a CXR-PRO dataset is proposed to deliberately exclude previous medical references to minimize the risk of hallucination. In addition, innovative evaluation techniques that account for and penalize hallucinations are also required, with these new metrics being designed to assess the accuracy and relevance of generated captions and their factual consistency with the input image and known medical knowledge [Yu et al.](#).

2.3 Conclusions

By critically analyzing the strengths and weaknesses of existing methods, this chapter identifies gaps in current research and lays the groundwork for developing advanced systems capable of addressing these challenges. The insights gathered here inform the methodological choices in subsequent chapters.

Chapter 3

Methodologies and Procedures

This chapter presents the core concepts that form the foundation of image captioning machine learning systems. A thorough understanding of these fundamental principles is essential to grasp the algorithms' complexities. Additionally, this chapter outlines the specific methodologies and experimental procedures employed throughout the research.

3.1 Main Methodology

Our approach leverages a lightweight architecture integrating pre-trained image encoders and text decoders to develop an effective medical image captioning model. The methodology also incorporates multiple labels to guide the caption generation process as additional semantic context. This work is structured into several key stages, each designed to explore different facets of the problem:

- **Baseline Analysis:** The initial phase involves evaluating state-of-the-art (SOTA) image captioning models from recent literature. To benchmark their performance, both zero-shot and few-shot evaluations will be performed on specific medical image datasets. The findings will help identify the most promising techniques for further development.
- **Encoder and Decoder Selection:** This stage focuses on experimenting with various image encoders and text decoder backbones to determine the most suitable. The selected ones will serve as the foundation for subsequent adaptations.
- **Image Captioning Enhancement:** The final phase focuses on refining the baseline captioning model through targeted adaptations. Multiple ablation studies will

be conducted to optimize performance. These experiments include experimenting with different configurations of the MIMIC-CXR dataset, investigating various methods to extract semantic information from images, and evaluating the effect of integrating medical labels on caption quality using three distinct fusion techniques.

The results from these stages will contribute to developing an enhanced medical image captioning model that effectively leverages domain-specific image labels to produce more accurate and contextually meaningful descriptions.

3.1.1 Image Captioning Workflow

Figure 3.1 depicts an overview of the model architecture for the standard image captioning workflow using the generative approach with mapping.

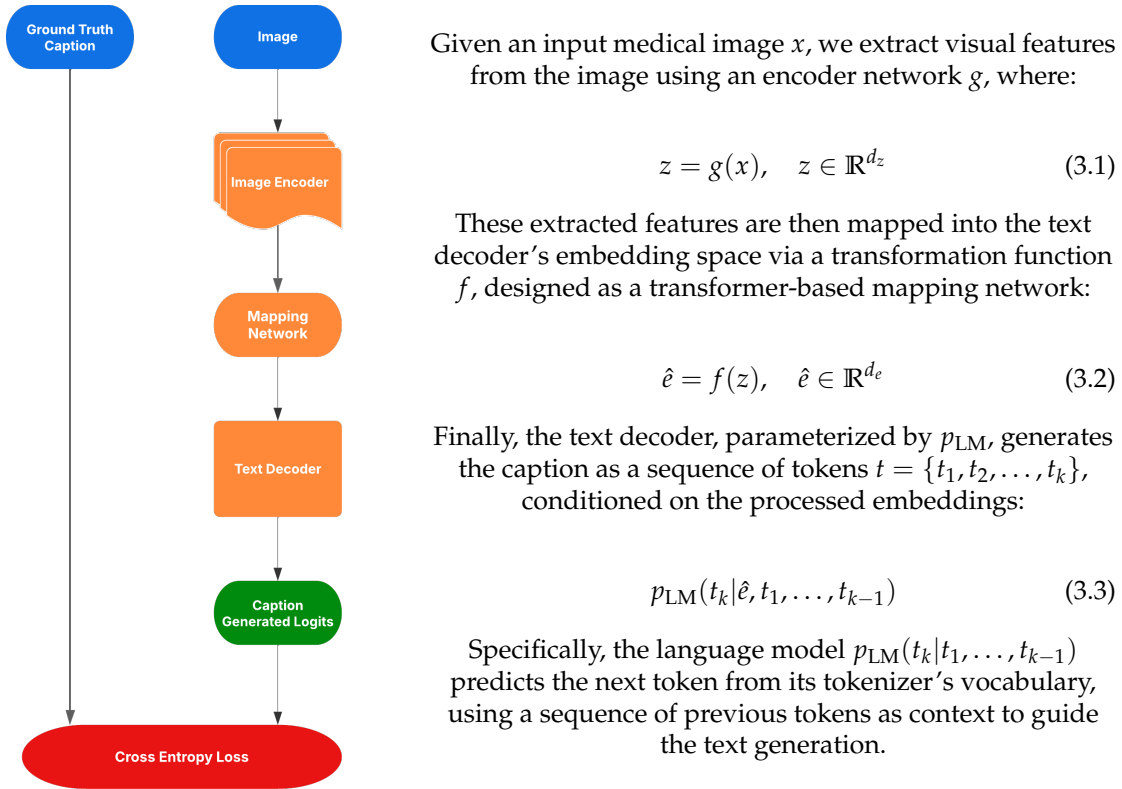


FIGURE 3.1: Standard Image Captioning Workflow

3.1.1.1 Image Encoders

There are two types of image encoders: Convolutional Neural Network (CNN) and Transformer-based image encoders.

CNNs became the game-changer proposed by [Lecun et al.](#). The proposed architecture, named LeNet-5, applies convolutional layers to understand the hierarchical pattern in data, a significant departure from fully connected layers as in Figure 3.2. They are widely used in applications such as image and video recognition, medical image analysis and Visual Question Answer (VQA).

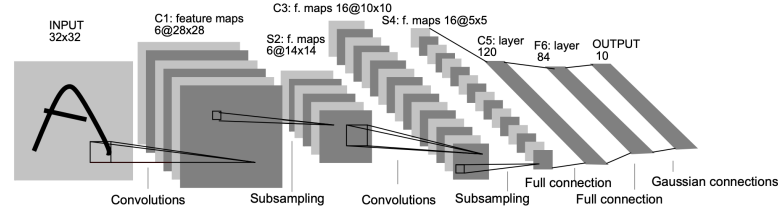


FIGURE 3.2: Architecture of LeNet-5 a Convolutional Neural Network for handwritten letters recognition [Lecun et al.](#).

In a CNN, hidden layers perform convolutions, which involve computing the dot product of a convolution kernel with the input matrix, followed by an activation function as depicted in Figure 3.3. As the kernel slides across the input, it generates feature maps contributing to the subsequent layer. This process reduces computation using shared weights across all image locations, making CNN more efficient than fully connected networks. However, the lack of interpretability is a serious challenge.

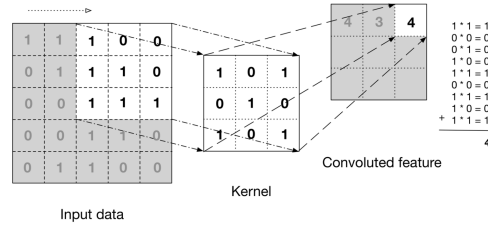


FIGURE 3.3: Computation of one step of the convolution [Roboflow](#).

Additional layers in CNN include **pooling layers**, which reduce data dimensions, **fully connected layers** connecting every neuron, and **normalization layers**. **Padding** preserves spatial resolution, and larger **stride** values reduce feature map dimensions. **Skip connections**, which bypass certain layers, facilitate the training of deep networks by improving optimization and mitigating issues like vanishing or exploding gradients.

Some examples of CNN based image encoders are: VGG [Simonyan and Zisserman](#), ResNet [He et al.](#), GoogleNet [Szegedy et al.](#), DenseNet [Huang et al.](#), EfficientNet [Tan and Le](#), MobileNet [Howard et al.](#) and Xception [Chollet](#).

The Transformer model introduced by Vaswani revolutionized Natural Language Pre-processing (NLP) and other domains by introducing self-attention mechanisms. The key innovation of the transformer is its ability to weight the importance of different elements in a sequence relative to each other, allowing the model to focus on the most relevant parts of the input.

The attention mechanism (Figure 3.4) relies on three components: **queries (Q)**, **keys (K)**, and **values (V)**. The input sequence of vectors, denoted as $\mathbf{X}$, is first projected into these three spaces using learned projection matrices:

$$\mathbf{Q} = \mathbf{XW}^{\mathbf{Q}}, \quad \mathbf{K} = \mathbf{XW}^{\mathbf{K}}, \quad \mathbf{V} = \mathbf{XW}^{\mathbf{V}}.$$

The core operation of the attention mechanism is the calculation of attention scores, which determine how much focus each element should receive. This is done using the formula:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{QK}^{\mathbf{T}}}{\sqrt{d_k}} \right) \mathbf{V}$$

where $\sqrt{d_k}$ is a scaling factor to prevent large values in the dot product, which could make training difficult by reducing gradient magnitudes.

Scaled Dot-Product Attention

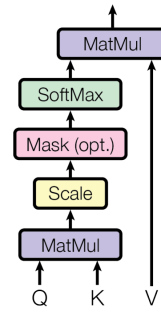
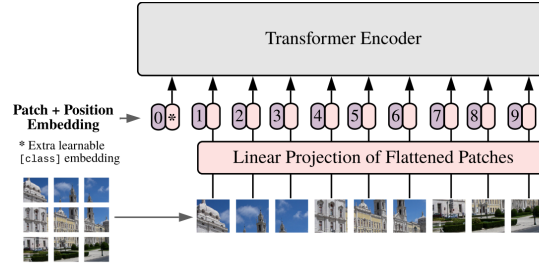


FIGURE 3.4: At-
tention Mechanism
Vaswani

Transformers are based on a multi-head attention mechanism and consist of two main components: the **encoder** and the **decoder**. The encoder processes the input sequence while the decoder generates the output sequence. Both are enhanced with positional encoding to capture the order of elements in the sequence. The transformer model was initially designed for machine translation but has since been applied to various domains, including computer vision, reinforcement learning, audio processing, and robotics.

Visual Transformers (ViT)s (Figure 3.5) extract patches from images and feed them into a Transformer encoder to obtain a global representation, which will finally be transformed for classification or other downstream tasks.

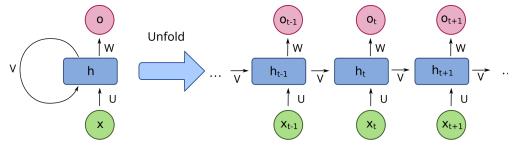
FIGURE 3.5: Vision Transformer [Dosovitskiy et al.](#)

Some examples of transformer-based image encoders are: ViT [Dosovitskiy et al.](#), Detection Transformer (DETR) [Carion et al.](#), SwinTransformer [Liu et al.](#), Cross Attention ViT (CrossViT) [Chen et al.](#) and Multiscale Vision Transformer (MViT) [Fan et al.](#)

3.1.1.2 Text Encoders and Decoders

For text generation, there are essentially two approaches: Recurrent Neural Network (RNN) and Transformer approaches.

RNNs are a class of artificial neural networks for sequential data processing. Unlike feedforward neural networks, which process data in a single pass, RNNs process data across multiple time steps, making them well-adapted for modelling and processing text, speech, and time series.

FIGURE 3.6: Recurrent Neural Network Architecture [neural network](#).

The recurrent unit is the fundamental building block of a RNN (Figure 3.6). It is a neural network that maps an input x_t into an output o_t , with the hidden vector h_t playing the role of "memory", a partial record of all previous input-output pairs. At each step, it transforms input into output and modifies its "memory" to help it perform better in future processing. This feedback loop allows the network to learn from past inputs and incorporate that knowledge into its current processing.

Early RNNs suffered from the vanishing gradient problem, limiting their ability to learn long-range dependencies. This gradient vanishing problem was solved by the invention of Long-Short Term Memory (LSTM) networks by [Hochreiter and Schmidhuber](#),

which set accuracy records in multiple application domains, becoming the default choice for RNN architecture.

In Text Transformers, text is converted to numerical representations called tokens, and each token is converted into a vector by looking up from a word embedding table. At each layer, each token is contextualized within the scope of the context window with other (unmasked) tokens via a parallel multi-head attention mechanism, allowing the signal for key tokens to be amplified and less important tokens to be diminished. It has led to developing pre-trained systems, such as Generative PreTrained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT).

Transformers (Figure 3.7) have the advantage of having no recurrent units, requiring less training time than earlier RNN such as LSTM.

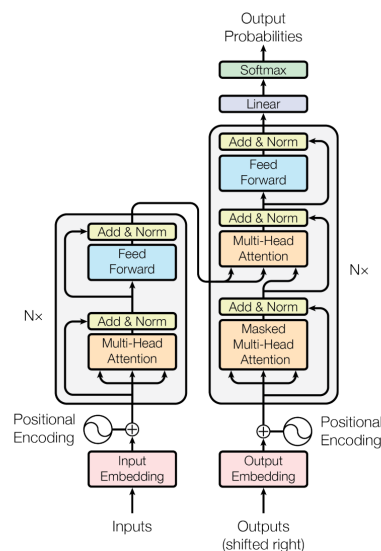


FIGURE 3.7: Text Transformer Architecture [Vaswani](#).

Some examples of transformer-based text decoders are GPT [Radford et al.](#) and T5 [Raffel et al.](#), while RNN comprises Google NMT [Wu et al.](#) and Seq2Seq LSTM [Sutskever et al.](#). Specifically to medical domains, the most relevant models are BioBert [Lee et al.](#) and MedGPT [Kraljevic et al.](#).

BioBert ([Lee et al.](#)) is a domain-specific language representation model built on top of BERT, specifically pre-trained on large-scale biomedical corpora. It is designed to better understand biomedical language and terminology compared to the original BERT model. BERT (Figure 3.8) uses the encoder-only transformer architecture [Devlin et al.](#) that learns to represent text as a sequence of vectors using self-supervised learning. Unlike earlier

models like GPT that read text left-to-right, BERT is trained in both directions, enabling a much richer understanding of context.

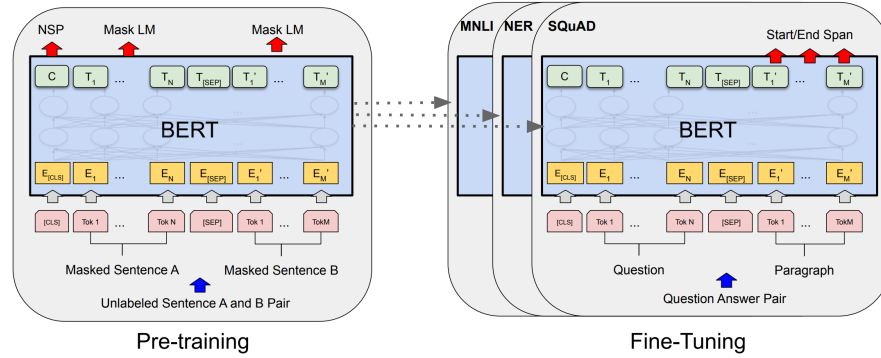


FIGURE 3.8: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in pre-training and fine-tuning. The same pre-trained model parameters are used to initialise models for different downstream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token. [Devlin et al.](#)

MedGPT ([Kraljevic et al.](#)) refers to a GPT model adapted or trained specifically for medical and clinical language tasks, similar to how BioBERT adapts BERT. GPT is an autoregressive language model developed by OpenAI ([Radford et al.](#)), based on the Transformer architecture (specifically, the decoder part). It is designed to generate human-like text by predicting the next word in a sequence, one token at a time. The Table 3.1 highlights the differences between BERT and GPT.

TABLE 3.1: Comparison between BERT and GPT models.

Feature	BERT	GPT
Directionality	Bidirectional	Unidirectional (left-to-right)
Architecture	Encoder only	Decoder only
Task Type	Understanding (e.g., classification)	Generation (e.g., text completion)
Pretraining Tasks	Masked Language Modeling (MLM) + Next Sentence Prediction (NSP)	Language Modeling (LM)

3.1.1.3 CLIP

[Radford et al.](#) proposed Contrastive Language-Image Pre-Training (CLIP) model. It is a combination of two powerful neural networks that are learned to connect images and text. It is trained on a large variety of image-text pairs using a contrastive learning approach, which enables it to understand visual concepts in a way that aligns closely with natural language. Unlike traditional models trained for specific tasks, CLIP (Figure 3.9) can generalize to a wide range of vision and language tasks without task-specific fine-tuning,

making it highly versatile for applications like zero-shot classification, image captioning, image and text retrieval and visual search, without requiring task-specific fine-tuning.

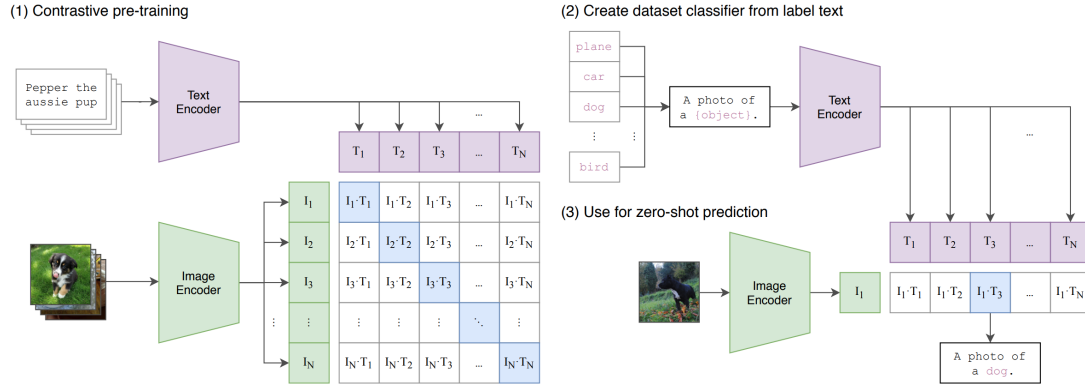


FIGURE 3.9: CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples(Radford et al.).

The main goal of this architecture is to encode both modalities into a common feature space where semantic similarities can be measured directly. During inference, CLIP assigns each image to the most closely aligned text by computing the cosine similarity between their respective feature vectors and selecting the pair with the highest score.

3.2 Datasets

During the pre-analysis phase of this study, both the Radiology Objects in COntext (ROCO) and MIMIC-CXR datasets will be used to explore and identify the most appropriate model architecture. Subsequently, the MIMIC-CXR dataset will serve as the primary training source for developing and refining the model's adaptation mechanisms. Finally, the NIH dataset will be used to evaluate the zero-shot prediction capabilities of the finalized model during the ablation study. Given their central role in this research, each dataset is briefly described below.

3.2.1 MIMIC-CXR Dataset

MIMIC-CXR-JPG v2.0.0 presented by Johnson et al. is a large dataset of 377,110 chest x-rays associated with 227,827 imaging studies sourced from the Beth Israel Deaconess Medical Center between 2011 - 2016.

The original raw captions (Johnson et al.) of the MIMIC-CXR dataset are semi-structured, with radiologists documenting their interpretations in titled sections such as "FINDINGS",

"IMPRESSION", "INDICATION", "REASON FOR EXAMINATION", "TECHNIQUE", "EXAMINATION", "COMPARISON" among others. The structure of these reports is generally consistent through standardized documentation templates. However, it can drift over time as the template changes or due to inter-reporter variability, as depicted in Figure 3.10.

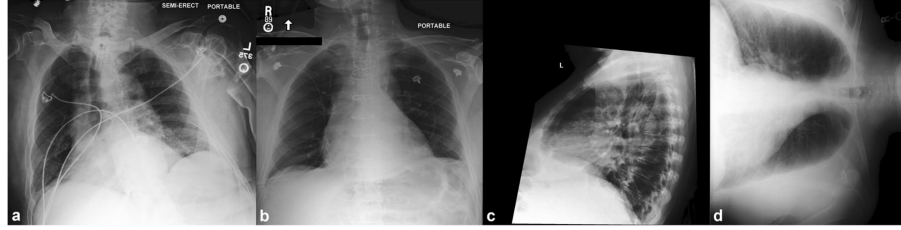


FIGURE 3.10: Images which highlight the amount of variation present in MIMIC-CXR.

Natural Language Explanation (NLE) captions were proposed by [Kayser et al.](#). Moreover, it aims to justify predictions made on medical images. NLE are human-friendly and comprehensive, enabling intrinsically explainable model training. Self-explanatory models have many benefits, including more robust and better-predicting performance, enabling the training of intrinsically interpretable models via a BERT-based explanation pipeline and extending the existing MIMIC-CXR dataset of chest X-rays with diagnoses, evidence labels, and NLE for the diagnoses as depicted in Figure 3.11.

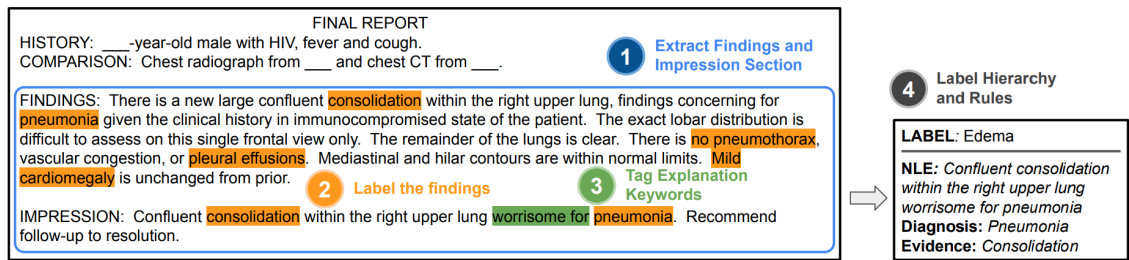


FIGURE 3.11: The steps required to extract NLEs from raw radiology reports. Findings and Impressions are extracted first, then labels refer to each sentence, generating sentences that contain explanation keywords, [Kayser et al.](#).

It was proposed to use self-explaining models that learn to detect lung conditions and explain their reasoning in natural language using a BERT-based labeller, targeted to clinical explanation keywords as depicted in Figure 3.12.

The learned image representations are constrained by mapping them to language that explains the evidence backing a diagnosis.

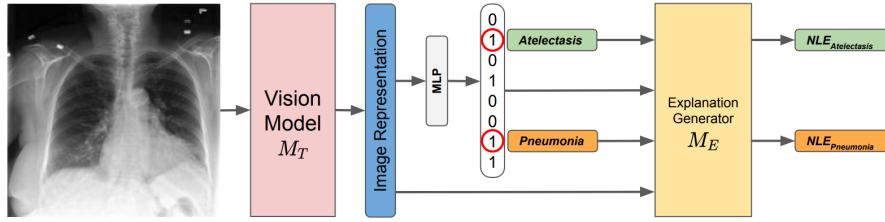
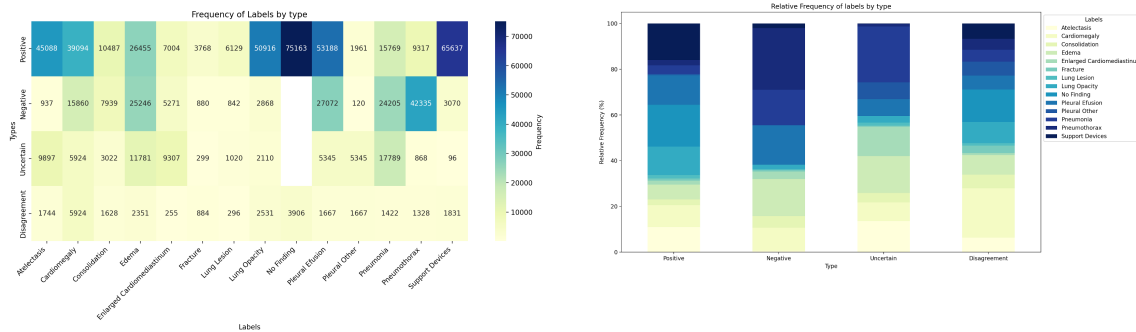


FIGURE 3.12: The model pipeline to provide an NLE for a prediction [Kayser et al.](#).

Labels in MIMIC-CXR dataset consists of 14 structured class categories derived from the database, with the distribution shown in Figure 3.13. All images are de-identified to protect patient privacy.



(A) Absolute Frequency of Labels by Type

(B) Relative Frequency of Labels by Type

FIGURE 3.13: Frequency of labels in MIMIC-CXR NLE on the training subset of 222,750 unique radiologic studies (8 studies were not labelled).

These labels align with the 14 structured categories used by CheXpert, which are employed to evaluate both the NegBio [Peng et al.](#) and CheXpert [Irvin et al.](#) labelers. The values are either blank (no value), 1.0 (mentioned and present), 0.0 (mentioned and absent), or -1.0 (mentioned and uncertain), with guidelines requiring annotation even for uncertain findings. Table 3.2 shows the dataset's training, validation, and testing sets distribution.

TABLE 3.2: Summary of the images split into training, validation, and test sets. [42].

Dataset	Train	Validate	Test
Number of Images	368960	2991	5159
Frontal	248020(67.2%)	2041(68.2%)	3653(70.8%)
Lateral	120795(32.7%)	949(31.7%)	1502(29.1%)
Other	145(0.0%)	1(0.0%)	4(0.1%)
Number of studies	222758	1808	3269
with a finding	170420(76.5%)	1394(77.1%)	2912(89.1%)
Number of patients	64586	500	293
with a finding	44157(68.4%)	344(68.8%)	288(98.3%)

Neither the CheXpert nor NegBio labellers are learnable classification models. They are rule-based systems that rely on NLP heuristics to extract labels from medical reports.

As such, they cannot be used as classifiers to compute a loss and perform backpropagation, as required by the Late Label Conditional Captioning framework (referred to in the Subsection 3.5). Therefore, they are not suitable for this approach. Instead, we can use pre-trained models such as ClinicalBERT or BioBERT, which are learnable and support gradient-based optimization.

3.2.2 ROCO Dataset

The ROCO dataset, presented by [Pelka et al.](#), is a multimodal dataset designed to explore the relationship between visual elements and semantic context in radiology images, curated by extracting image-caption pairs from the open-access biomedical literature database, PubMedCentral [92].

The ROCO dataset consists of over 81k radiology images across multiple modalities, including Computed Tomography (CT), Ultrasound, X-ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), and Angiography. Each image is paired with a caption, keywords, Unified Medical Language System (UMLS), Concept Unique Identifiers (CUIs), and Semantic Types (SemTypes).

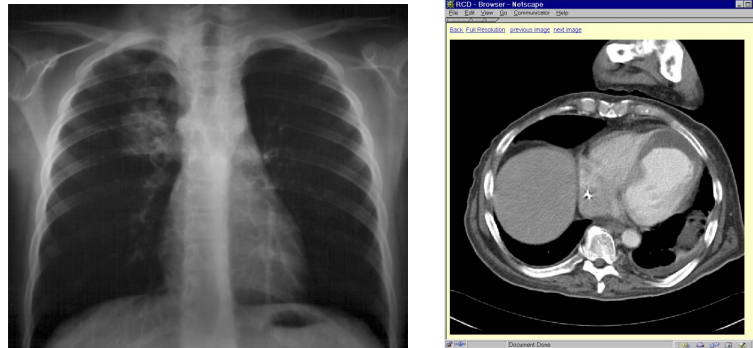


FIGURE 3.14: Examples of ROCO images and respective captions: **Left** - "Chest radiograph. Parenchymal density in the right upper lobe." **Right** - "Detailed view of a single image in the teaching file. Using the navigation bar, all images of a case can be viewed sequentially in high resolution."

The dataset is split into two main classes: Radiology and Non-Radiology. The Radiology class contains 81,825 images, while the Non-Radiology class has 6,127 images, split into training (80%), validation (10%), and test (10%) sets. Figure 3.15 illustrates the workflow used for the dataset's development.

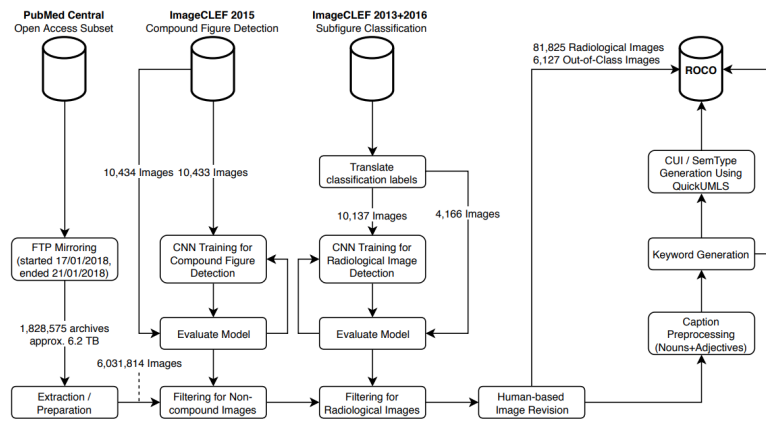


FIGURE 3.15: Overview of the workflow used for the ROCO dataset development [Pelka et al.](#).

3.2.3 NIH Dataset

The NIH Chest X-ray Dataset (Figure 3.16) consists of an extensive collection of chest radiographs (X-rays) collected from patients at the Indiana University Hospital and other associated institutions. It is designed to support developing and evaluating medical image analysis algorithms, particularly for automated disease detection and report generation. This dataset includes over 100,000 frontal-view chest X-ray images from more than 30,000 unique patients and associated radiology reports, which include detailed findings and impressions written by radiologists.

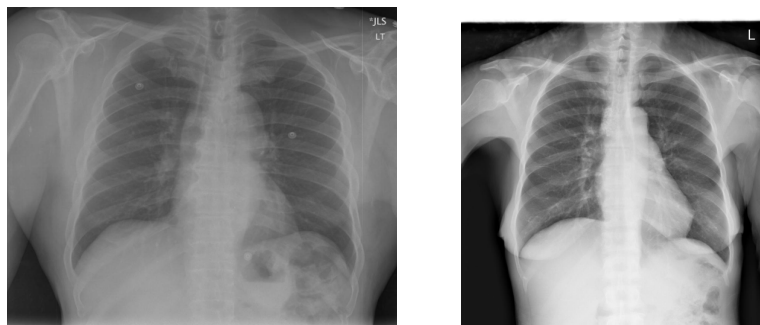


FIGURE 3.16: Examples of NIH images and respective captions: **Left** - "XXXX-year-old XXXX with pain. Cardiomedastinal silhouette demonstrates normal heart size with tortuosity and atherosclerosis of the thoracic aorta. No focal consolidation, pneumothorax, or pleural effusion. No acute bony abnormality identified. Multilevel degenerative disc disease of the thoracic spine was noted. No acute cardiopulmonary abnormality."; **Right** - "Chest x-XXXX dated XXXX at XXXX hours. XXXX-year-old female, transplant workup. Calcified mediastinal XXXX. No focal areas of consolidation. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Degenerative changes thoracic spine. No acute cardiopulmonary abnormality."

Similar to the MIMIC-CXR dataset, it can be extracted 14 labels using NegBio or CheXpert labelers.

3.3 Evaluation and Performance Metrics

Evaluating multimodal medical image captioning systems requires both linguistic and clinical accuracy. Standard metrics from natural language processing are used to assess syntactic and semantic quality, while domain-specific metrics help determine clinical relevance.

In order to evaluate the quality of the generated captions, the metrics used by [Yu et al.](#) were applied. It examined the correlation between 31 automated metrics and the scoring of reports by radiologists. They constructed metric-oracle reports for four metrics that include BiLingual Evaluation Understudy (BLEU), BERTScore, CheXbert vector similarity, RadGraph23 F1 and Radiology Report Clinical Quality (RadCliQ), by selecting the highest scoring report, according to each of the metrics, from the training set (Figure 3.17).

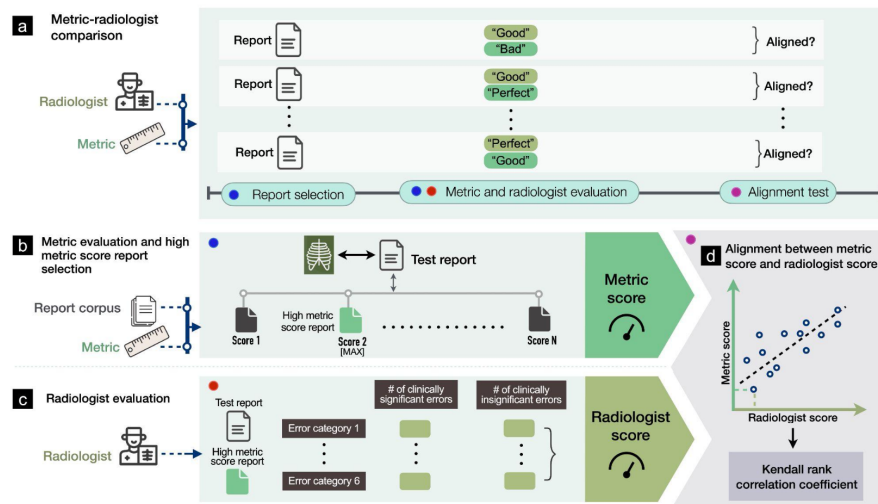


FIGURE 3.17: a) Experimental design for selecting radiology reports and comparing metrics and radiologists in evaluating reports. b) Given a test report, select the report with the highest metric score from the training report corpus with respect to the test report and a particular metric. c) Conducting radiologist evaluation on the high metric score report relative to the test report, where radiologists identify the number of clinically significant and insignificant errors in the high metric score report across six error categories. d) Determining the alignment between metric scores and radiologists scores assigned to the same reports using the Kendall rank correlation coefficient. [Yu et al.](#)

Then, six board-certified radiologists scored the number of errors that various metric oracle reports make compared to the test report. Radiologists categorized errors as significant or insignificant. Radiologists subtyped every error into six categories. They first quantify metric-radiologist alignment using the Kendall rank correlation coefficient between metric scores and the number of radiologist-reported errors in the reports. In addition to evaluating the clinical relevance of metrics in terms of the total number of

clinically significant and insignificant errors, the particular error categories of metric oracles were also examined to develop a granular understanding of the failure modes of different metrics.

Besides the metrics presented by Yu et al., other traditional metrics like Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) were also computed.

3.3.1 Domain Specific Metrics

RadCliQ can rank the quality of reports similarly (Composite Score). The goal is to minimize this metric, which means a lower error relative to human perception.

RadGraph, on the other hand, is a metric that computes the overlap in clinical entities and relations between a machine-generated report and a radiologist-generated report. Overall, **RadGraph** was found to be the best individual metric to use for its strong alignment with radiologists, reliability across failure modes and meaningful empirical performance with generations. RadGraph F1 score measures the F1-score between medical entities and extracted relation by RadGraph and varies between 0 and 1.

CheXbert vector similarity is designed to measure the correctness of clinical information. It computes the cosine similarity between the indicator vectors of 14 pathologies that the CheXbert automatic labeller extracts from machine-generated and human-generated radiology reports. So, it varies between 0 and 1 and is designed to evaluate radiology-specific information, but its evaluation is limited to 14 pathologies. To address this limitation, Yu et al. proposes the use of the knowledge graph of the report to represent arbitrarily diverse radiology-specific information.

3.3.2 Tradicional NLP Metrics

BLEU score evaluates how similar a generated text (hypothesis) is to a reference text by comparing their n-grams (consecutive sequences of words). It ranges from 0 to 1, where a higher score indicates a closer match. Mathematically, the BLEU score is defined as:

$$BLEU = \min \left(1, \exp \left(1 - \frac{\text{reference-length}}{\text{output-length}} \right) \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}} \quad (3.4)$$

where:

$$\text{precision}_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m_{\text{cand}}^i, m_{\text{ref}}^i)}{\sum_{\text{snt}' \in \text{Cand-Corpus}} \sum_{i' \in \text{snt}'} m_{\text{cand}}^{i'}} \quad (3.5)$$

with:

- m_{cand}^i is the count of i-gram in candidate matching the reference translation.
- m_{ref}^i is the count of i-gram in the reference translation.
- w_t^i is the total number of i-grams in the candidate translation.

ROUGE evaluates how well the generated text overlaps with the reference text. It ranges from 0 to 1, with higher values indicating better similarity. It includes several variants:

- **ROUGE-1:** Measures the overlap of single words (unigrams) between the generated and reference texts. It is a basic measure of lexical similarity.
- **ROUGE-2:** Measures the overlap of word pairs (bigrams), indicating how well keyword sequences are preserved.
- **ROUGE-L:** Focuses on the Longest Common Subsequence (LCS), assessing structural and syntactic similarity. This is particularly useful in medical text evaluation, as it captures the alignment of terminology and structured information more effectively than simple word matching.

Unlike BLEU and ROUGE, which focus on exact word matches, the generic **BERTScore** leverages contextual embeddings from the generic BERT model to measure the semantic similarity between the generated and reference texts. It ranges from 0 to 1, with higher values indicating better alignment in meaning.

BERT score presented by [Zhang et al.](#) in 2020, consists of three components:

- **Precision:** Measures how well the words in the generated text align with the reference text.
- **Recall:** Measures how well the generated text covers the words in the reference text.
- **F1-Score:** The harmonic mean of precision and recall offers a balanced evaluation.

METEOR improves upon BLEU by considering more than exact word matches. It incorporates:

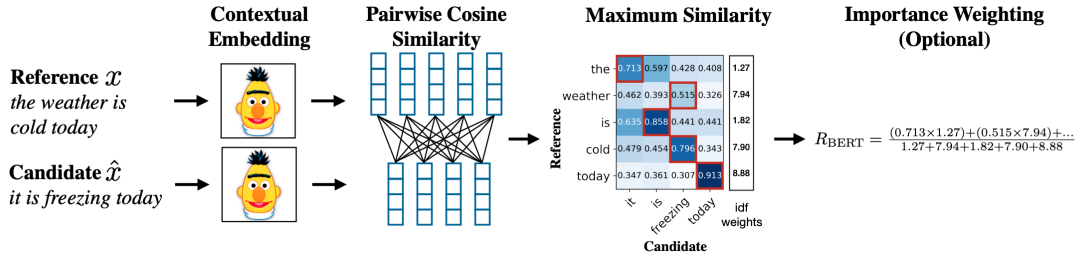


FIGURE 3.18: Illustration of the computation of the recall metric RBERT. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red and include the optional importance weighting [Zhang et al.](#).

- **Synonym matching:** recognizing different words with similar meanings.
- **Stemming:** matching root forms of words, such as "run" and "running".
- **Word order alignment:** ensuring that phrases maintain logical structure. This makes METEOR particularly useful for domains like medical text generation, where synonyms and terminology variations are common.

CIDEr is primarily used for evaluating caption generation, considering the best individual metric. It ranges from 0 to 10, where 0 means no similarity to the reference captions and 10 indicates near-perfect similarity. Real-world evaluations often yield values between 0 and 3. CIDEr assigns higher importance to uncommon but relevant words using term frequency-inverse document frequency (TF-IDF), ensuring that rare but meaningful terms significantly impact the score. It was first presented in 2014 by [Vedantam et al.](#). The CIDEr_n score for n -grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|}, \quad (3.6)$$

where $\mathbf{g}^n(c_i)$ is a vector formed by $g_k(c_i)$ corresponding to all n -grams of length n , and $\|\mathbf{g}^n(c_i)\|$ is its magnitude. Similarly, for $\mathbf{g}^n(s_{ij})$. To account for grammatical properties and semantics, we combine the scores from varying n -gram lengths as follows:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i), \quad (3.7)$$

where empirically found that uniform weights $w_n = \frac{1}{N}$ work best. We use $N = 4$.

3.3.3 Overview

Each of these metrics serves a unique purpose. While BLEU and ROUGE measure lexical overlap and help evaluate structured text, BERTScore assesses semantic similarity, making it effective for tasks requiring deeper meaning alignment. CIDEr score is tailored for image captioning and weighs the importance of specific words, andxw METEOR provides a balanced evaluation by considering synonyms, stemming, and word order. Combining these metrics allows one to obtain a more comprehensive evaluation of text generation quality.

3.4 Baseline Models

This section overviews the baseline models used in developing the image captioning pipeline. These models serve as points of comparison and sources of inspiration, with their most promising components subsequently integrated into the final captioning architecture. To gauge performance, each baseline will be evaluated using zero-shot and few-shot learning approaches on the MIMIC-CXR NLE and ROCO datasets. This evaluation aims to identify the strengths and weaknesses of each model, guiding the refinement and enhancement of the captioning system. The resulting insights will directly inform subsequent pipeline optimisation and integration phases.

3.4.1 ClipCap Model

ClipCap ([Mokady et al.](#)) extends the CLIP model, designed for image-text retrieval tasks, to the image captioning domain. The proposed architecture is depicted in Figure 3.19.

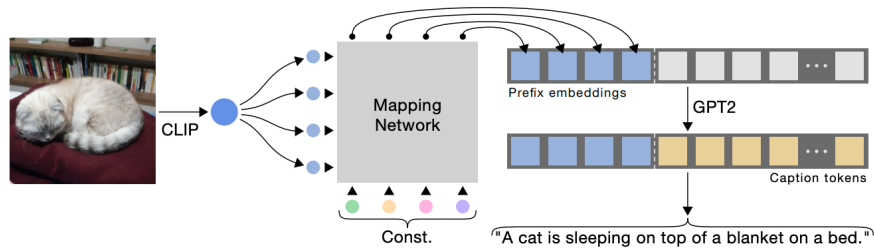


FIGURE 3.19: Overview of the transformer-based architecture, enabling the generation of meaningful captions while both CLIP and the language model, GPT-2, are frozen. We train a lightweight transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2 to extract a fixed-length prefix. It was employed GPT-2 at inference to generate the caption given the prefix embeddings. Also, suggest an MLP-based architecture ([Mokady et al.](#)).

The approach involves extracting image features using an image encoder, which are then passed through a mapping network to project these features into the text decoder's feature space. Those are then used as prefixes for the caption generator. The mapping network can be implemented using a transformer-based or Multilayer Perceptron (MLP) based architecture. Additionally, text generation methods, such as Greedy Search and Beam Search, generate captions, offering different strategies for selecting the most probable caption sequences.

3.4.2 BLIP Model

Li et al. presents BLIP, a unified model for both generation tasks (image captioning and VQA) and retrieval tasks (image text retrieval) as depicted in Figure 3.20.

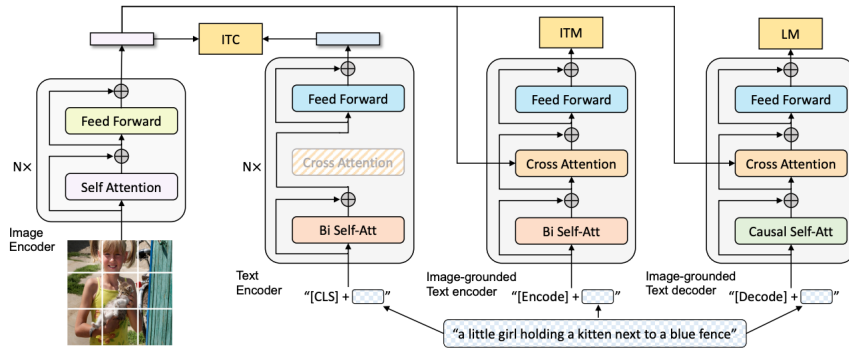


FIGURE 3.20: Pre-training model architecture and objectives of BLIP (same parameters have the same color). Propose a multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions and is trained with an image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images (Li et al.)

This model introduces two contributions:

- Multimodal mixture of Encoder-Decoder, MED: a model for effective multi-task pre-training. It can operate as an unimodal encoder, image-grounded text encoder, or image-grounded text decoder.
- Captioning and Filtering, CapFilt: a data bootstrapping method for learning from noisy text-image pairs. MED was fine-tuned into two modules: a captioner to produce synthetic captions from web images and a filter to remove noisy captions from the original and synthetic texts.

3.4.3 MediCap Model

[Nicolson et al.](#) proposes MediCap, an image captioning method that leverages an encoder-decoder model combined with reinforcement learning to optimize the model based on a primary evaluation metric, in this case, the BERT score. The method explores two approaches: one uses cross-attention to merge the image features and prompt, which are then passed to the text decoder, while the other uses self-attention, where image features are mapped into the text decoder's feature space and serve as a prompt. The training process consists of two stages: Teacher Forcing (TF) and Self-Critical Sequence Training (SCST), allowing for enhanced captioning performance through both supervised and reinforcement learning strategies (Figure 3.21).

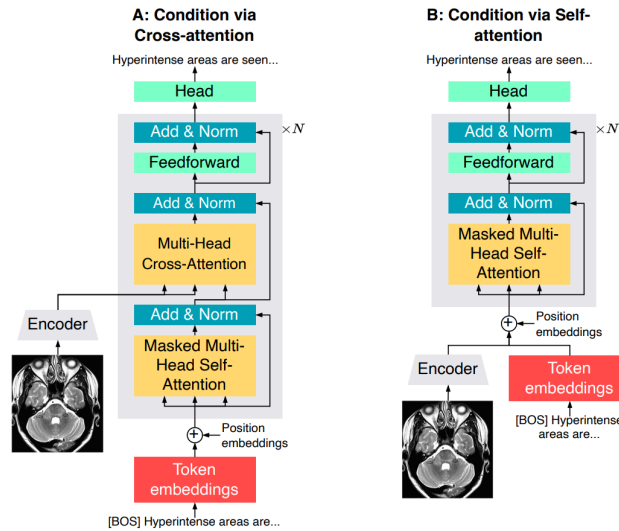


FIGURE 3.21: Decoder conditioned on the visual features of the image via A) the cross-attention, and B) the self-attention. The visual features are extracted with the encoder. N is the number of Transformer blocks. [BOS] is the special token for the beginning of the sentence ([Nicolson et al.](#)).

3.4.4 DiCO Model

[Moratelli et al.](#) propose DiCO (Direct CLIP-Based Optimization), a new training paradigm designed to align image captioning models with human preferences while preserving fluency and stability. DiCO directly optimises a learnable captioning evaluator alongside the captioner by solving a weighted classification problem within the framework. This approach prevents divergence from the original model and maintains high linguistic quality in the generated captions. DiCO shows improved stability and enhanced caption quality and aligns more closely with human preferences, outperforming existing methods across modern metrics. The architecture of DiCO is illustrated in Figure 3.22.

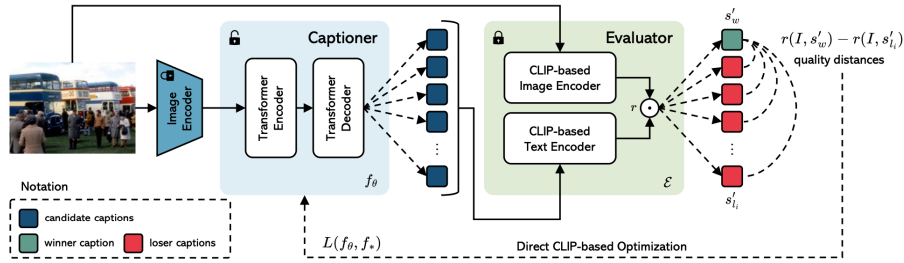


FIGURE 3.22: Given an image and candidate generations, the figure shows the process for captioner fine-tuning by distilling from a CLIP-based evaluator. ([Moratelli et al.](#)).

One key feature that sets DiCO apart from related methods, such as SCST, is its ability to distil a reward model directly from a contrastive-based evaluator without employing Reinforcement Learning during fine-tuning. Furthermore, DiCO generates caption candidates directly from the model, removing the need for human-annotated preference datasets, often a bottleneck in Reinforcement Learning from Human Feedback methods. DiCO also includes a regularizer to keep the fine-tuned model close to its pre-trained counterpart, thereby avoiding catastrophic drift while improving its performance.

3.5 Label Conditioned Image Captioning

To include the labels in the captioning process, we have three approaches to follow:

- Early Label Conditioned Caption Generation/Learnable Prompt Caption Generation.
- Intermediate Label Conditioned Caption Generation.
- Late Label Conditioned Caption Generation.

Intermediate Label Conditioning Image Captioning (Figure 3.23) refers to a model that uses the labels in a soft prompt fixed by "The image shows: (...)". The soft prompt is then encoded by the text encoder chosen, and the image is encoded by the corresponding image encoder and projected into the text decoder feature space by a transformer-based mapping network. Both image and label embeddings, projected in the text decoder embedding space, are concatenated into a final text decoder embedding array by a cross-attention mapper. The output then feeds the text decoder to generate the caption.

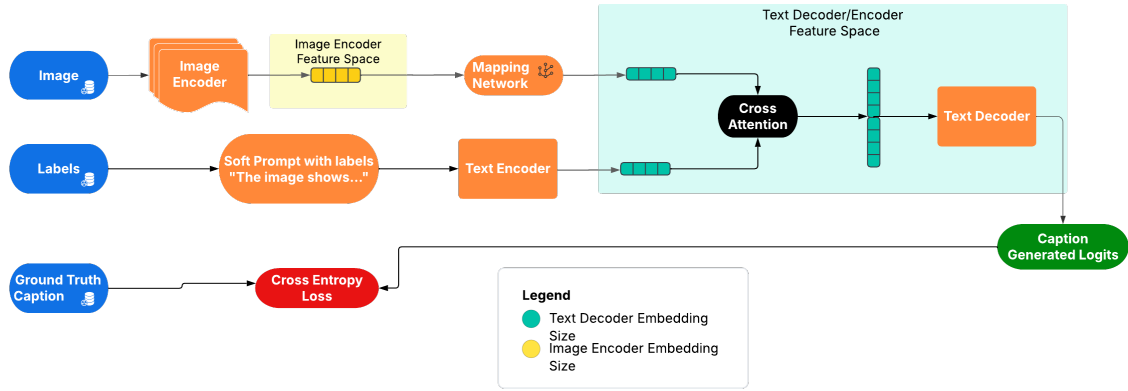


FIGURE 3.23: Intermediate Label Conditioned Captioning: Architecture

Model Description

Given a medical image $\mathbf{x}$, a classifier $h(\mathbf{x})$ predicts a set of labels $\mathcal{L} = \{l_1, l_2, \dots, l_m\}$, which are incorporated into a soft prompt of the form:

$$\text{"The image shows: } l_1, l_2, \dots, l_m \text{"}$$

This prompt is encoded into a label embedding using the same language model p_{LM} that will later be used for caption generation, resulting in a label embedding $\mathbf{p} \in \mathbb{R}^{d_p}$.

Visual features are extracted from the image using an image encoder g , producing:

$$\mathbf{z} = g(\mathbf{x}), \quad \mathbf{z} \in \mathbb{R}^{d_z}$$

These visual features are then projected into the text decoder embedding space via a transformer-based mapping function f :

$$\hat{\mathbf{e}} = f(\mathbf{z}), \quad \hat{\mathbf{e}} \in \mathbb{R}^{d_e}$$

To integrate the label semantics into the visual representation, cross-attention is applied, where the projected image embedding is used as the query, and the label embedding serves as the key and value:

$$\hat{\mathbf{u}} = \text{CrossAttention}(\text{query} = \hat{\mathbf{e}}, \text{key} = \mathbf{p}, \text{value} = \mathbf{p}), \quad \hat{\mathbf{u}} \in \mathbb{R}^{d_e}$$

This fused embedding $\hat{\mathbf{u}}$ is used as input to the language model p_{LM} , which generates a caption token by token:

$$p_{\text{LM}}(t_k | \hat{\mathbf{u}}, t_1, \dots, t_{k-1})$$

The inclusion of cross attention was inspired by BLIP model (Li et al.) while the base captioner architecture was based on ClipCap (Mokady et al.).

Early Label Conditioned Image Captioning (Figure 3.24) refers to an approach similar to the previous one. However, instead of having a fixed soft prompt, it uses the hot labels and creates a label embedding through a Label Mapping Network. Then, the label and image encoding are concatenated, as previously mentioned.

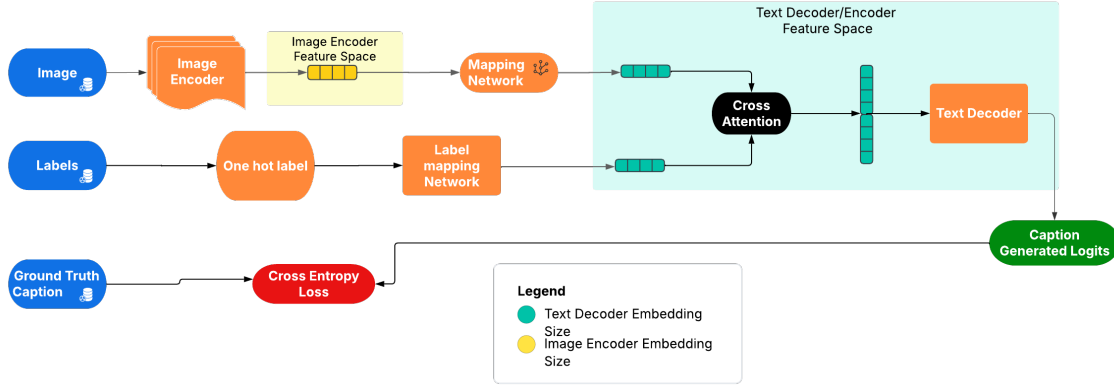


FIGURE 3.24: Early Label Conditioned Captioning: Architecture

Model Description

Given a medical image $\mathbf{x}$, a classifier $h(\mathbf{x})$ predicts a set of labels $\mathcal{L} = \{l_1, l_2, \dots, l_m\}$ in one-hot representation, which are mapped into a learnable label embedding through a label mapping network, $\mathbf{p} \in \mathbb{R}^{d_p}$.

Then, visual features are extracted from the image using an image encoder g , such that $\mathbf{z} = g(\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^{d_z}$. The visual features are then projected into the text decoder embedding space via a transformer-based mapping f : $\hat{\mathbf{e}} = f(\mathbf{z})$, with $\hat{\mathbf{e}} \in \mathbb{R}^{d_e}$.

To integrate the label semantics into the visual representation, cross-attention is applied, where the projected image embedding is used as the query, and the label embedding serves as the key and value:

$$\hat{\mathbf{u}} = \text{CrossAttention}(\text{query} = \hat{\mathbf{e}}, \text{key} = \mathbf{p}, \text{value} = \mathbf{p}),$$

resulting in the fused representation $\hat{\mathbf{u}} \in \mathbb{R}^{d_e}$.

This fused embedding is used as input to a language model p_{LM} , which generates a caption token by token:

$$p_{\text{LM}}(t_k \mid \hat{\mathbf{u}}, t_1, \dots, t_{k-1}).$$

Lastly, **Late Label Conditioned Image Captioning** (Figure 3.25) refers to an image captioning approach that only uses the labels in the loss calculation. The idea is to use the standard image captioning architecture where the image features are projected into the text decoder feature space, generating the caption. Then, the classifier receives the predicted caption as input to predict the multilabel vector. In this case, the loss component is not only associated with the captions (cross-entropy loss). However, it is also affected by a term corresponding to the Binary Cross Entropy loss calculated between the ground truth labels and the predicted labels from the generated caption.

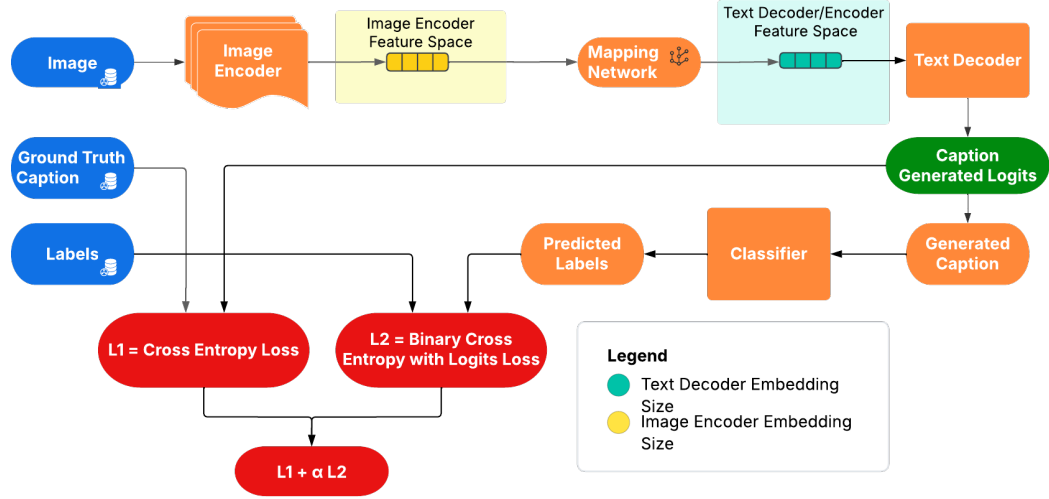


FIGURE 3.25: Late Label Conditioned Captioning: Architecture

Model Description

Given a medical caption $\mathbf{y}$, a classifier $\langle(\mathbf{y})$ predicts a set of labels

$$\mathcal{L} = \{\ell_1, \ell_2, \dots, \ell_m\}$$

Features are extracted from an image $\mathbf{x}$ using an image encoder g , producing

$$\mathbf{z} = g(\mathbf{x}), \quad \mathbf{z} \in \mathbb{R}^{d_z}$$

The visual features are then projected into the text decoder embedding space via a transformer-based mapping f :

$$\hat{\mathbf{e}} = f(\mathbf{z}), \quad \hat{\mathbf{e}} \in \mathbb{R}^{d_e}$$

A language model p_{LM} generates a caption token by token:

$$p_{\text{LM}}(t_k \mid \hat{\mathbf{e}}, t_1, \dots, t_{k-1})$$

The predicted caption $\hat{\mathbf{y}}$ is then passed through the classifier $p(\hat{\mathbf{y}})$ to get the predicted set of labels:

$$\hat{\mathcal{L}} = \{\hat{\ell}_1, \hat{\ell}_2, \dots, \hat{\ell}_n\}$$

The final loss is defined as the weighted sum of the cross-entropy loss associated with the predicted and ground-truth captions, and the binary cross-entropy loss associated with the predicted and ground-truth set of labels:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_2 \cdot \mathcal{L}_{\text{BCE}}(\hat{\mathcal{L}}, \mathcal{L})$$

3.6 Training Methods

It is worth pausing to reflect on the multiple training methods (Table 3.3).

Autoregressive Conventional Models, like ClipCap, already referred to, are a standard autoregressive approach where the model generates text step by step based on previously predicted tokens. This approach has the advantage that the model learns to generate sequences naturally by predicting the next word based on its outputs. However, errors accumulate over time since the model relies on its imperfect predictions.

Teacher Forcing is a method in which the model is trained using the actual ground truth sequence rather than its previous predictions. This strategy allows faster and more stable training, learning more efficiently and converging quickly. However, during inference (when generating text), the model does not have access to the ground truth and must rely on its predictions, which can lead to errors if it has not learned to handle mistakes.

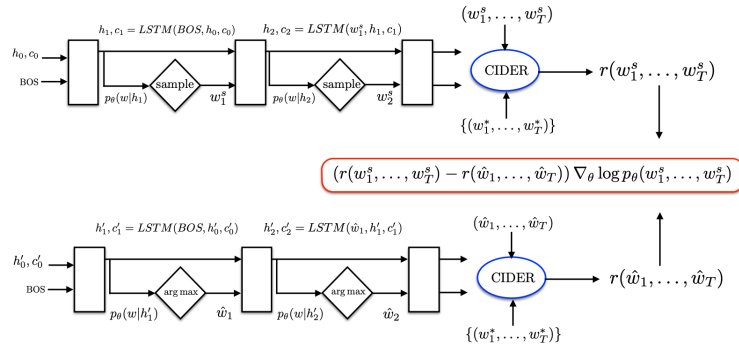


FIGURE 3.26: Self-critical sequence training (SCST). The weight put on words of a sampled sentence from the model is determined by the difference between the reward for the sampled sentence and the reward obtained by the estimated sentence under the test-time inference procedure (greedy inference depicted). This harmonises learning with the inference procedure and lowers the variance of the gradients, improving the training procedure, (Rennie et al.).

Lastly, **Self-Critical Sequence Training** is a reinforcement learning (RL) method designed to improve text generation models using a reward-based approach to optimize model predictions. This approach encourages the model to explore better alternatives instead of relying on deterministic methods and optimizes outputs based on meaningful semantic similarity rather than word-by-word matching. It also reduces over-reliance on greedy search, leading to more diverse and natural text generation. This method was presented for image captioning by Rennie et al. in 2016 with the architecture in Figure 3.26.

Table 3.3 summarizes the training methods available.

TABLE 3.3: Comparison of training methods and their architectures.

Training Method	Architecture / Process Description
Autoregressive Conventional Models	For example, if the ground truth is "The cat is sleeping" the model does the following: • Step 1: The model receives [start] and predicts "The". • Step 2: The model uses its own prediction "The" as input to predict "cat". • Step 3: This continues token-by-token until the sentence is complete.
Teacher Forcing	For example, if the ground truth is "The cat is sleeping" the model does the following: • Step 1: The model receives [start] and predicts "The". • Step 2: Instead of using its own prediction, the correct token "The" is fed to predict "cat". • Step 3: This continues with the ground truth tokens used at each step (e.g., "cat", "is", ...).
Self-Critical Sequence Training	It works the following way: • The model generates two captions: – Sampled (stochastic): e.g., "A dog runs on the grass." – Greedy (deterministic): e.g., "Happy dog running park." • Compute semantic similarity (e.g., with BERTScore) to reference caption. • Use the reward difference (e.g., $0.85 - 0.78 = 0.07$) to compute the loss. • Update the model to increase the likelihood of high-reward sequences.

3.7 Text Generation Methods

Two main text generation methods are often used, mainly Beam search and Greedy search.

The generation process differences are shown in Figure 3.27.

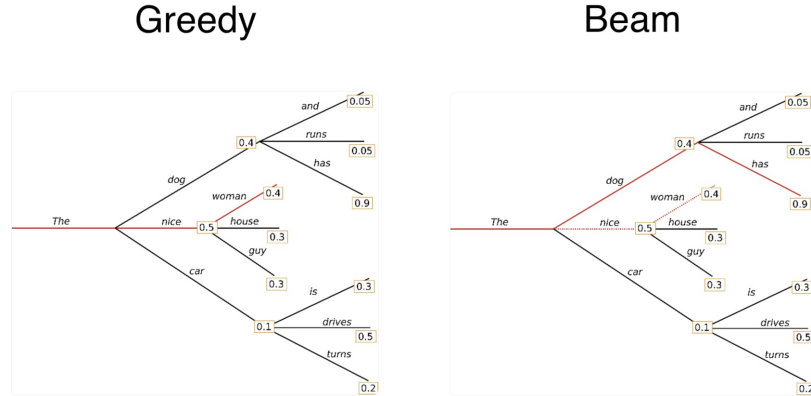


FIGURE 3.27: Beam and Greedy Search Methods

Greedy Search generates text by always selecting the word with the highest probability at each step. While this approach is efficient and fast, it may produce sub-optimal or repetitive output due to its limited consideration of future possibilities.

Beam search maintains multiple candidate sequences (determined by a beam width) and expands them in parallel. Exploring multiple paths before choosing the best candidate helps produce more coherent and fluent text. However, it can be computationally

intensive and may still miss the best possible output if it settles on a local optimum. **Sampling search** introduces randomness by sampling words from the probability distribution instead of choosing the most likely one. This sampling allows for greater creativity and variation in the generated text, although it sometimes produces less coherent or nonsensical output.

3.8 Conclusions

In this chapter, we presented a comprehensive overview of the methodologies employed in this study. We described the range of image captioning approaches, including baseline models and the subsequent adaptations implemented to enhance performance, mainly label conditional captioning.

We also described the datasets used and outlined the evaluation metrics to quantitatively gauge the models' output, offering a clear framework for comparing traditional and domain-specific metrics. Furthermore, the training procedures were discussed, encompassing the optimization strategies and text generation mechanisms that enable the models to produce coherent and contextually accurate captions.

Together, these components provide a strong foundation for the subsequent experimental work, allowing for a rigorous and insightful assessment of the models' capabilities in generating high-calibre image captions.

Chapter 4

Experiments and Results

This chapter presents the experimental evaluation of various models, beginning with baseline approaches drawn from the literature. These models, originally developed for general image captioning tasks, are assessed in both zero-shot and few-shot settings within the medical domain. As they are pre-trained on large-scale, general-purpose datasets, they serve as initial reference points to gauge generalization capabilities prior to any domain-specific adaptation. Building upon the insights from these baselines, we then introduce a series of targeted modifications to a selected baseline model. These enhancements include the integration of different pre-trained image encoder and text decoder backbones, the use of diverse captions adaptations, the incorporation of auxiliary structured labels across three modalities, and additional architectural refinements.

The chapter culminates with the development of a fully trained, lightweight medical image captioning model that surpasses the performance of existing state-of-the-art (SOTA) approaches in this domain.

4.1 Performance Criteria and Model Selection

The workflow of this research is structured as in Figure 4.1: We begin with zero-shot and few-shot training of baseline image captioning architectures, as described in Chapter 3. This initial phase enables the identification of the strengths and limitations of each model, which will inform subsequent improvements to the final architecture. Next, we evaluate a variety of image encoders and text decoders to select the most promising combinations to refine the final model further. These subsequent experiments involve a series of training adaptations, including variations in training hyperparameters, caption type selection,

and the incorporation of reinforcement learning techniques such as Self-Critical Sequence Training (SCST) and other architectural enhancements.

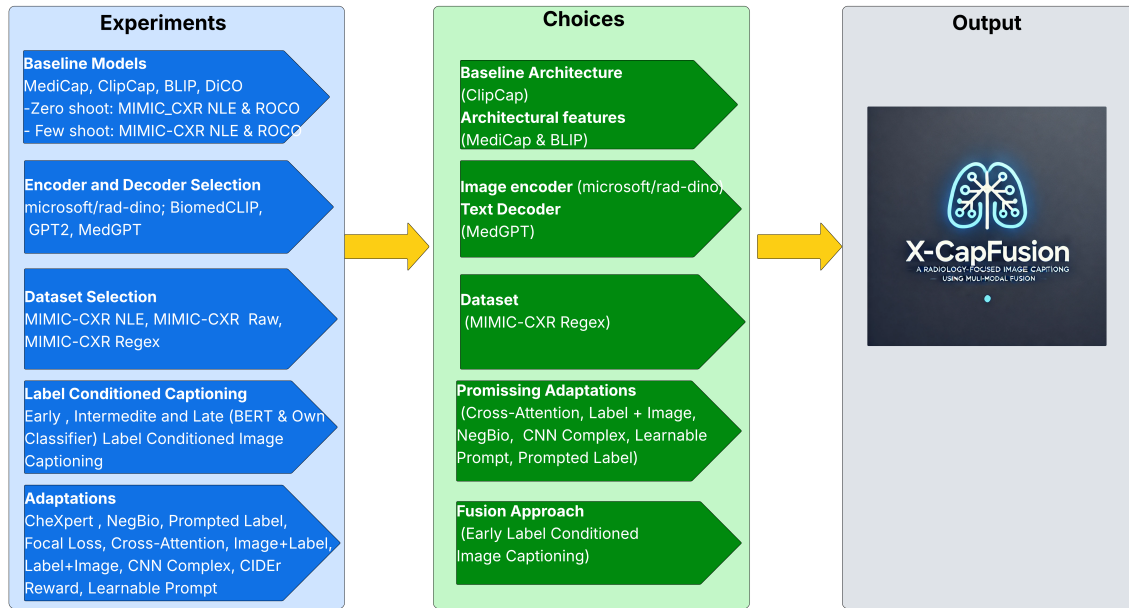


FIGURE 4.1: X-CapFusion Research Workflow

The most effective adaptations are then carried forward into the final phase, which involves the integration of semantic labels in label-conditioned captioning models using three conditioning strategies: early, intermediate, and late-label integration frameworks.

Ultimately, a unified model—*X-CapFusion*—is developed by combining the best-performing components and techniques. This model is then rigorously evaluated and compared against baseline models through a comprehensive ablation study.

Each model variation is assessed using a wide range of both traditional and domain-specific evaluation metrics, including: BiLingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Bidirectional Encoder Representations from Transformers (BERT)Score, Consensus-based Image Description Evaluation (CIDEr), Sem-bScore, RadGraph Combined, RadCliQ-v0 and RadCliQ-v1. This multi-faceted evaluation thoroughly assesses caption quality from both linguistic and clinical perspectives.

4.2 Baseline Models

This first section presents the analysis of two specific datasets of some of the SOTA image captioning models already described in Section 3.4. The idea is to make zero-shot and

few-shot predictions in MIMIC-CXR Natural Language Explanation (NLE) and ROCO datasets. The parameters used for prediction in zero-shot and training in few-shot scenarios are shown in Appendix A.

4.2.1 Zero shot Results

For zero shot evaluation in **ClipCap** it was used ViT-B-32 as the image encoder and GPT2 as the text decoder with a transformer-based mapping network layer, using the checkpoint proposed in Mokady et al. using either COCO and Conceptual captions dataset. On the other hand, **BLIP** used ViT B-16 as image encoder and BERT as text decoder and the corresponding checkpoint is made available by Li et al.. **MediCap** used CvT-21 as image encoder and DistilGPT2 as text decoder and was training either by Teacher forcing and SCST (Nicolson et al.). Finally, **DiCO** used ViT-L/14 as an image encoder and a designed transformer as a text decoder (Moratelli et al.).

Table 4.1 shows the prediction metrics obtained for zero-shot training in the Radiology Objects in Context (ROCO) and MIMIC-CXR NLE datasets of the image captioning baseline pre-trained models referred to above.

TABLE 4.1: Baseline Models: Zero-shot Prediction Metrics in ROCO and MIMIC-CXR NLE datasets

Metric	ClipCap		BLIP	DiCO	MediCap
	COCO	Conceptual Captions			
ROUGE1 Precision	0.068	0.079	0.112	0.079	0.238
ROUGE1 Recall	0.126	0.212	0.201	0.137	0.316
ROUGE1 F1	0.080	0.104	0.131	0.091	0.248
ROUGE2 Precision	0.003	0.013	0.017	0.003	0.072
ROUGE2 Recall	0.006	0.048	0.030	0.007	0.099
ROUGE2 F1	0.003	0.019	0.020	0.004	0.076
ROUGEL Precision	0.063	0.072	0.098	0.072	0.208
ROUGEL Recall	0.114	0.192	0.172	0.121	0.270
ROUGEL F1	0.074	0.095	0.113	0.082	0.215
BERT Precision	0.811	0.812	0.818	0.805	0.857
BERT Recall	0.858	0.869	0.857	0.838	0.883
BERT F1	0.834	0.839	0.837	0.821	0.870
METEOR	0.059	0.094	0.093	0.063	0.183
CIDEr	0.007	0.016	0.026	0.007	0.113
BLEU [127]	0.014	0.020	0.022	0.017	0.079
SembScore [127]	0.617	0.609	0.633	0.623	0.683
RadGraph Combined [127]	0.001	0.002	0.004	0.002	0.043
RadCliQ-v0 [127]	4.151	3.897	3.961	4.176	3.232
RadCliQ-v1 [127]	1.716	1.574	1.607	1.729	1.173

(a) ROCO Dataset

Metric	ClipCap		BLIP	DiCO	MediCap
	COCO	Conceptual Captions			
ROUGE1 Precision	0.062	0.062	0.056	0.073	0.071
ROUGE1 Recall	0.087	0.118	0.100	0.084	0.108
ROUGE1 F1	0.070	0.077	0.070	0.075	0.083
ROUGE2 Precision	0.001	0.003	0.001	0.007	0.010
ROUGE2 Recall	0.001	0.006	0.002	0.007	0.013
ROUGE2 F1	0.001	0.004	0.001	0.007	0.011
ROUGEL Precision	0.057	0.055	0.051	0.065	0.064
ROUGEL Recall	0.079	0.103	0.090	0.074	0.096
ROUGEL F1	0.064	0.068	0.064	0.066	0.074
BERT Precision	0.827	0.828	0.829	0.821	0.849
BERT Recall	0.846	0.868	0.848	0.824	0.863
BERT F1	0.836	0.848	0.838	0.823	0.855
METEOR	0.043	0.057	0.044	0.047	0.049
CIDEr	0.012	0.011	0.012	0.009	0.038
BLEU [127]	0.024	0.028	0.012	0.028	0.037
SembScore [127]	0.344	0.325	0.355	0.392	0.525
RadGraph Combined [127]	0.001	0.001	0.000	0.000	0.049
RadCliQ-v0 [127]	4.646	4.591	4.482	4.660	3.713
RadCliQ-v1 [127]	1.974	1.942	1.890	1.984	1.446

(b)MIMIC NLE Dataset

MediCap consistently outperformed all other models in zero-shot evaluations on the ROCO dataset. Although both the ClipCap model pre-trained on Conceptual Captions

and BLIP showed strong performance on the MIMIC-NLE dataset, overall model performance was notably higher on ROCO than on MIMIC-NLE. This discrepancy may be attributed to the repetitive and lexically limited nature of the MIMIC-NLE reference captions, which often rely heavily on recurring medical terms such as “pulmonary” and “atelectasis”. These repetitive patterns can lead to lower recall scores, particularly in models that generate more diverse or paraphrased outputs, since recall depends on exact word matches.

Notably, the ClipCap model pre-trained on Conceptual Captions consistently outperformed its counterpart trained on the COCO dataset across all metrics. This performance gain is likely due to the larger and more diverse nature of the Conceptual Captions dataset, which supports better generalization than the more domain-specific and limited scope of COCO, which primarily focuses on everyday images. As a result, only the Conceptual Captions checkpoint is used in subsequent few-shot experimental setups.

In particular, both the DiCO and ClipCap models trained on COCO exhibited notably low CIDEr scores on the ROCO dataset. This suggests a significant divergence between the generated captions and the human references, reinforcing the limitation of COCO-trained models in medical domains.

Finally, MediCap’s superior performance—especially in generating lexically diverse captions—suggests that it may have been trained on a more comprehensive and heterogeneous medical dataset compared to MIMIC-NLE. This diversity likely contributed to its reduced reliance on repetitive phrasing and improved generalization in clinical image captioning tasks.

Images 4.2 and 4.3 show predictions made from the baseline models in a zero-shot setting using a random example of ROCO and MIMIC-CXR NLE test set, encoded with a scheme of color codes *.

MediCap provides a caption that matches the modality and diagnosis, well aligned with the medical domain description in the ROCO dataset. However, it does not find the correct location for the MIMIC-CXR NLE example from a patient’s point of view. It might have been trained in a dataset where the locations (right and left) were seen from a doctor’s point of view (reverse).

*Green means the concept is presented in the caption and the image; Red means the concept is not by the caption and image; Blue means it is not presented in the caption but can be seen from the image or is inconclusive.

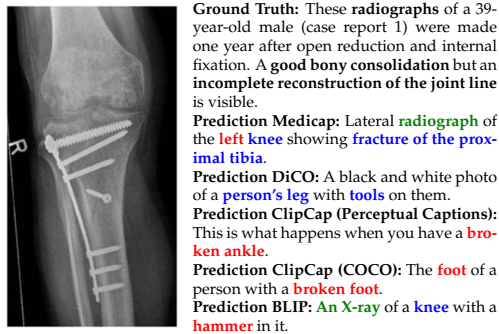


FIGURE 4.2: Predictions for a ROCO test set example.

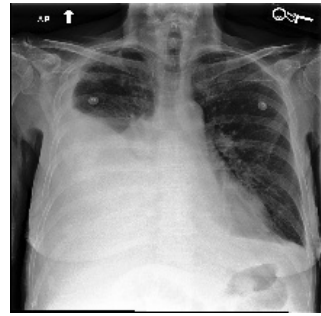


FIGURE 4.3: Predictions for a MIMIC NLE test set example.

DiCO model correctly identifies grayscale medical images and the presence of a leg but fails to provide clinical or diagnostic details present in ROCO dataset and completely hallucinates with the MIMIC-CXR NLE example. It also shows signs of overfitting and generates repetitive predictions, especially when trained on limited data. These issues might be alleviated through dataset expansion or text augmentation techniques.

ClipCap predictions present a colloquial or formal text tone, with redundant phrasing, no modality recognition and incorrect anatomy identification (ankle instead of knee), lacking clinical correctness and image grounding. Although correct, BLIP shows some hallucination due to limited domain exposure in the ROCO dataset and a very general description in the MIMIC-CXR NLE example.

This performance confirms that strong medical-domain pretraining is essential in zero-shot scenarios, and general captioners are unreliable in clinical settings. Models pre-trained on general datasets struggle to generate accurate, relevant, and diverse captions in medical contexts, often producing repetitive or generic outputs, contrasting with the MediCap model. As it was trained on a medical dataset, it has a larger understanding of medical terminology, relationships, and concepts, allowing it to handle unseen medical images accurately.

4.2.2 Few shot Predictions

For a few-shot training experiments in MIMIC NLE dataset in **MediCap**, the model was trained for 100 epochs, and captions were generated using Beam Search methodology with 5 beams. On the other hand, **BLIP** was trained shortly for 5 epochs, using nucleus sampling as a text generation method. **MediCap** was training for 10 epochs (Teacher forcing) and 3 epochs (SCST), using Beam Search as a text generation method with a

beam size of 4. Finally, **DiCO** was trained by 10000 steps (without reward) and 250000 steps (with reward) and used Beam Search as text generation methodology.

Table 4.2 shows the results obtained for a few-shot training on ROCO and MIMIC-CXR NLE datasets.

TABLE 4.2: Baseline Models: Few-shot Prediction Metrics in ROCO and MIMIC-CXR NLE datasets

Metric	ClipCap (COCO)	BLIP	DiCO	MediCap	Metric	ClipCap (COCO)	BLIP	DiCO	MediCap
ROUGE1 Precision	0.176	0.211	0.084	0.224	ROUGE1 Precision	0.190	0.210	0.116	0.177
ROUGE1 Recall	0.230	0.327	0.150	0.289	ROUGE1 Recall	0.186	0.257	0.123	0.248
ROUGE1 F1	0.175	0.233	0.096	0.232	ROUGE1 F1	0.178	0.222	0.114	0.201
ROUGE2 Precision	0.047	0.067	0.008	0.063	ROUGE2 Precision	0.045	0.069	0.012	0.056
ROUGE2 Recall	0.063	0.108	0.013	0.083	ROUGE2 Recall	0.043	0.086	0.012	0.085
ROUGE2 F1	0.046	0.074	0.009	0.065	ROUGE2 F1	0.046	0.073	0.011	0.066
ROUGEL Precision	0.156	0.187	0.080	0.193	ROUGEL Precision	0.163	0.184	0.103	0.164
ROUGEL Recall	0.202	0.285	0.143	0.243	ROUGEL Recall	0.159	0.224	0.108	0.227
ROUGEL F1	0.154	0.205	0.092	0.197	ROUGEL F1	0.152	0.194	0.101	0.185
BERT Precision	0.845	0.842	0.801	0.855	BERT Precision	0.868	0.857	0.834	0.881
BERT Recall	0.853	0.869	0.804	0.880	BERT Recall	0.865	0.866	0.809	0.885
BERT F1	0.849	0.855	0.802	0.867	BERT F1	0.866	0.861	0.821	0.883
METEOR	0.120	0.178	0.066	0.1667	METEOR	0.109	0.165	0.064	0.158
CIDEr	0.108	0.116	0.023	0.082	CIDEr	0.144	0.195	0.073	0.017
BLEU [127]	0.155	0.064	0.018	0.082	BLEU [127]	0.070	0.081	0.029	0.096
SembScore [127]	0.330	0.698	0.684	0.685	SembScore [127]	0.484	0.553	0.492	0.536
RadGraph Combined [127]	0.157	0.038	0.000	0.028	RadGraph Combined [127]	0.072	0.107	0.066	0.102
RadCliQ-v0 [127]	4.036	3.385	3.954	3.311	RadCliQ-v0 [127]	3.649	3.461	4.020	3.372
RadCliQ-v1 [127]	1.530	1.264	1.609	1.221	RadCliQ-v1 [127]	1.388	1.270	1.610	1.216

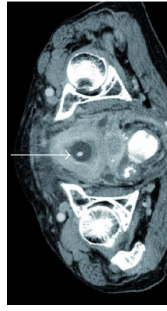
(a) ROCO Dataset
(b) MIMIC-CXR NLE Dataset

MediCap and BLIP show superior performance in all metrics in a few-shot scenario, followed by the ClipCap model.

DiCO performed poorly in all metrics, especially in generating syntactically meaningful captions. Its struggle to capture relevant content is evident in low ROUGE and CIDEr scores, indicating its inability to produce diverse and meaningful descriptions.

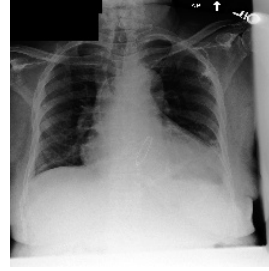
ClipCap achieved also a good performance being a very lightweight and simple, when compared with the others. Figures 4.4 and 4.5 show predictions made from baseline models in a few shot setting on a specific example of ROCO and MIMIC-CXR NLE test set.

MediCap and ClipCap tend to produce more syntactically coherent, medically plausible descriptions. DiCO generates noisy or redundant output, often repeating words (e.g., “aspect aspect aspect”, “probably probably”). BLIP often provides semantically close descriptions but misses specific or mislocates anatomical details (e.g., mass in rectum vs. bladder wall).



Ground Truth: Axial computed tomography scan of the pelvis showing a diffuse infiltration of the bladder wall, catheter in situ (arrow).
 Prediction MedicaP: **Computed tomography** scan of the **abdomen** showing a **calcification in the left pelvic artery** (arrow).
 Prediction DiCO: **CT view of gland** of al aspect (**arrow** aspect aspect aspect).
 Prediction ClipCap: **Computed tomography** of the **pelvis** revealed **thickened bladder wall**.
 Prediction BLIP: **CT** scan of the **pelvis** showing a **mass in the right rectum**.

FIGURE 4.4: Predictions for a ROCO test set example.



Ground Truth: There is persistent minimal blunting of the left costophrenic angle, which may represent a tiny effusion or chronic pleural thickening.
 Prediction MedicaP: **Chest X-ray** showing a **left pleural effusion**.
 Prediction DiCO: **Right** lung opacity acity lung probably probably **ateateateleclis**.
 Prediction ClipCap: Increased interstitial markings are seen throughout the lungs suggesting **pulmonary vascular congestion**.
 Prediction BLIP: There is mild prominence of the pulmonary interstitial markings suggestive of **mild pulmonary**.

FIGURE 4.5: Predictions for a MIMIC NLE test set example.

Due to its simplicity, lightweight architecture, and competitive performance, we selected the ClipCap base model as the foundation for further adaptations. However, to enhance its multimodal capabilities, we will incorporate several techniques inspired by both BLIP and MediCap models.

4.3 Encoder and Decoder Selection

This section will train models using various combinations of image and text decoders by freezing the encoder and decoder models and only training the mapping networks, as shown in Figure 4.6. This approach allows us to focus on the zero-shot capabilities of the encoder and decoder models for our specific medical dataset, enabling the selection of the most promising encoder and facilitating the training of the final image captioning model. The mapping network that will be used is based on the one used by ClipCap model by Mokady et al..

Table 4.7 displays the parameters used for training these models with multiple image encoders and text decoders. Table 4.3 also shows the choices of image and text decoder pairs used in each experiment.

TABLE 4.3: Image Encoders and Text Decoders Pairs

Experiment	Image Encoder	Text Decoder
Rad+GPT2 Experiment	microsoft/rad-dino [82]	gpt-2 [85]
Rad+medGPT Experiment	microsoft/rad-dino [82]	Rabbiaaa/MedGPT [46]
BioCLIP+GPT2 Experiment	hf-hub:microsoft/BiomedCLIP-PubMedBERT [129]	gpt-2 [85]
ViT+GPT2 Experiment	ViT-B/32 [87]	gpt-2 [85]

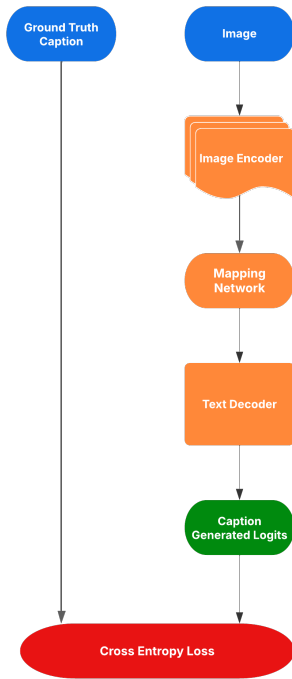


FIGURE 4.6: Archi-
tecture

FIGURE 4.7: Training Parameters for Text De-
coder and Image Encoder Selection

Parameter	Description
Dataset	MIMIC-CXR NLE
Batch size	16
Image Embedding Prefix Length	10
Transformer Mapping	Transformer Based (8 layers)
Learning rate	2e-5
Warmup steps	5000
Prefix Length Clip	10
Train Image Encoder	False
Train Mapping	True
Train Text Decoder	False
Loss	Cross Entropy Loss
Epochs	20
Text Generation	Beam Search
Number beams	5
Generation Length	60
Temperature	1.0
Stop token	''

Figure 4.8 shows the loss evolution over 20 epochs for each model experiment. In the loss history analysis, experiments with Rad+GPT2 and Rad+medGPT exhibit better convergence than the other experiments. While the remaining experiments show rapid convergence, their final loss values remain relatively high. Notably, the Experiment with Rad+medGPT resulted in a lower final loss than the Experiment with Rad+GPT2.

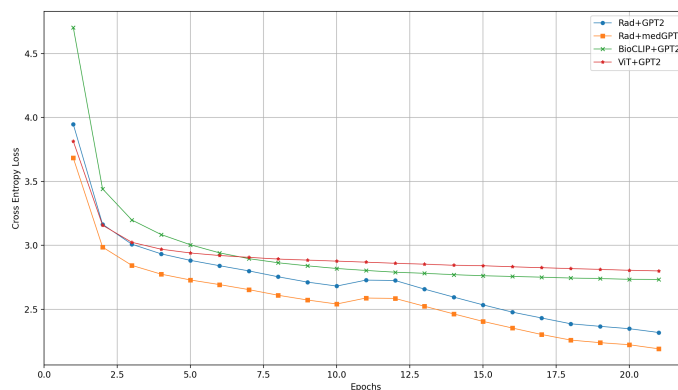


FIGURE 4.8: Image Encoder and Text Decoder Selection Experiments: Epoch Loss
Overview

Table 4.4 displays the prediction metrics obtained in the experiments. Regarding evaluation metrics, some favour Experiment Rad+GPT2, while others show a slight advantage for Experiment Rad+medGPT, with a few metrics also favouring Experiment BioCLIP+GPT2. Experiment ViT+GPT2 represents the original ClipCap model, which serves as a baseline for comparison (Mokady et al.). that consistently shows a drop in most metrics compared with the other experiments.

TABLE 4.4: Image Encoder and Text Decoder Selection Experiments on MIMIC-CXR NLE Dataset.

Metric	Rad+GPT2	Rad+medGPT	BioCLIP+GPT2	ViT+GPT2
ROUGE1 Precision	0.202	0.187	0.176	0.144
ROUGE1 Recall	0.286	0.294	0.123	0.272
ROUGE1 F1	0.225	0.212	0.209	0.181
ROUGE2 Precision	0.061	0.056	0.012	0.035
ROUGE2 Recall	0.088	0.091	0.012	0.068
ROUGE2 F1	0.068	0.065	0.066	0.044
ROUGEL Precision	0.184	0.166	0.103	0.137
ROUGEL Recall	0.260	0.280	0.108	0.259
ROUGEL F1	0.205	0.197	0.198	0.172
BERT Precision	0.873	0.871	0.867	0.863
BERT Recall	0.880	0.876	0.868	0.865
BERT F1	0.876	0.873	0.867	0.863
METEOR	0.177	0.166	0.170	0.142
BLEU [127]	0.098	0.094	0.091	0.069
SembScore [127]	0.553	0.549	0.563	0.549
RadGraph Combined [127]	0.122	0.116	0.117	0.092
RadCliQ-v0 [127]	3.307	3.358	3.370	3.467
RadCliQ-v1 [127]	1.173	1.204	1.212	1.282

Experiment Rad+GPT2 consistently scores the best on ROUGE1, ROUGE2, and ROUGEL F1 and precision, indicating overall caption quality and balance. However, ROUGE recall is better for Experiment Rad+medGPT, indicating maximal content coverage at the cost of some precision. Experiment BioCLIP+GPT2 has a very low recall for ROUGE1 and ROUGE2, suggesting it generates shorter or less detailed outputs. RadGraph Combined is highest in Experiment Rad+GPT2, indicating better clinical content correctness.

Experiment Rad+medGPT is close to Experiment Rad+GPT2 in surface-level overlap, indicating that it produces captions structurally similar to the references text and maintains good semantic alignment when compared with ground truth captions, exhibiting a high BERT Score. It is a well-balanced model, maintaining linguistic quality and semantic relevance but lacking clinical precision, with a slight advantage that Experiment Rad+GPT2 provides. However, given that Experiment Rad+medGPT employs image encoders and text decoders pre-trained on medical datasets, it is expected to perform better in few-shot training scenarios (mainly while fine tuning the text decoder). It is important to note that the evaluation was conducted with both the text decoder and text encoder

frozen.

Figure 4.9 presents predictions on a specific test set example of MIMIC-CXR NLE dataset.

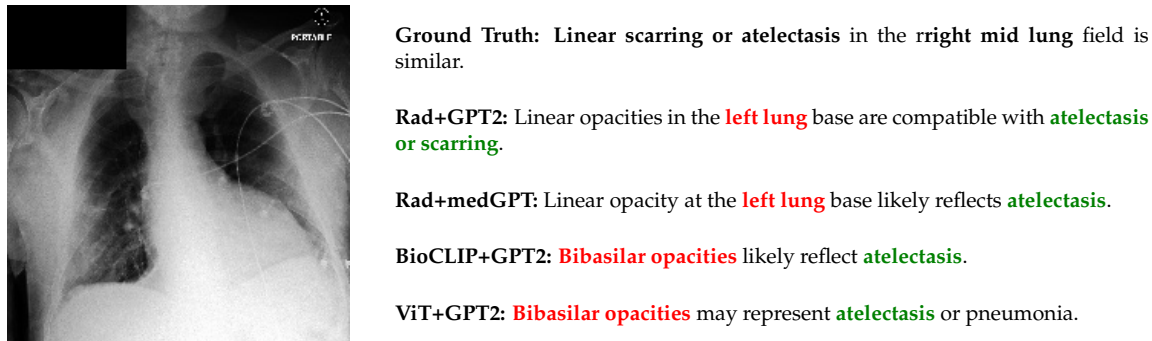


FIGURE 4.9: Predictions for a MIMIC-CXR NLE test set example.

In the Rad+GPT2 and Rad+medGPT experiments, although the models do not consistently identify the correct anatomical side—likely due to ambiguity in perspective (e.g., doctor’s vs. patient’s point of view)—they successfully detect the presence of the relevant anomaly. In contrast, BioCLIP+GPT2 and ViT+GPT2 models fail to identify the anomaly in the same scenarios.

This limitation is partly attributable to the nature of the MIMIC-NLE dataset, where models trained on such data often overuse frequent terms like “atelectasis” and “pulmonary”, reflecting a bias in the training corpus. These repetitive patterns can be mitigated by applying data augmentation techniques such as back-translation and paraphrasing, which introduce greater lexical and syntactic diversity.

It was also noted that in the early stages of training, models frequently generate near-identical captions across test images—a phenomenon known as initial bias. As training progresses, fine-tuning of the model’s weights helps to overcome this issue and improve output quality. Additionally, low temperature settings during caption generation can further reduce output diversity by making the model overly deterministic, which may hinder its ability to produce varied and nuanced descriptions.

Finally, smaller batch sizes, while allowing for more frequent parameter updates, introduce noisier gradients, especially noticeable in the early training epochs. This can affect the stability and convergence rate of the model but may also promote generalization under certain conditions.

4.3.1 Visual Explanations with Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al.) is a widely used technique for visualizing the most relevant regions of an image that contribute to a model’s decision. It works by computing the gradients of a specific class output concerning the activations of a given convolutional layer, typically the last one, and combining these gradients to produce a class-discriminative heatmap. This heatmap highlights the areas in the input image that are most influential in making a prediction.

In image captioning, Grad-CAM can provide valuable insights into how the model interprets visual content to generate descriptive text. Specifically, when using a dual-stage architecture — where a frozen image encoder is used to extract image features, and a text decoder generates captions - Grad-CAM can help visualize which parts of the image contribute most to specific words or phrases in the generated caption.

In order to guide in the selection of the best image encoder, GradCam results were analyzed for the three pre-trained image encoders used as depicted in Figure 4.10 for a specific example. Note that Experiments Rad+GPT2 and Rad+medGPT use the same encoder frozen, so only one example represents both.

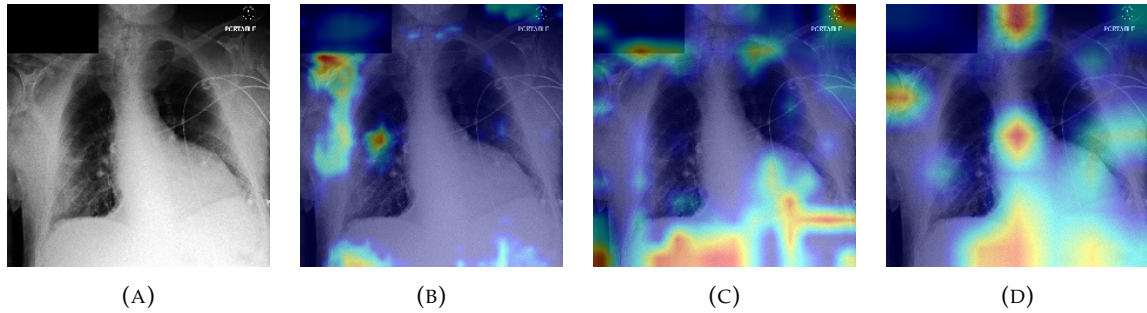


FIGURE 4.10: GradCam: Analysis of the last layer of the image encoders. (a) Image with caption: “Linear scarring or atelectasis in the right mid lung field is similar.” (b) Microsoft Rad-DINO image encoder: layer `model.encoder.layer.11.output`. (c) Microsoft BioMedCLIP image encoder: layer `visual.trunk.blocks.4`. (d) ViT-B/32 image encoder: layer `clip_model.visual.transformer.resblocks.11`.

According to medical opinion, the Image that displays less noise is Image A. Additionally, Image A partially detects an opacity/atelectasis in the right mid-lung, which is reflected in the ground truth caption. Image C slightly identifies the right mid-lung opacity but has too much noise, giving too much importance to regions of the Image that are not clinically relevant to the caption. Image B also displays a lot of noise and does not correctly detect the opacity/atelectasis in the right mid-lung.

The fact that Image A is more specific and has less noise suggests that the model can be fine-tuned to achieve better results than the others, allowing focus on relevant parts of the medical Image.

So, based on all these results, further experiments will be conducted using Experiment Rad+medGPT configuration, including Rabbiaaa/MedGPT for the text decoder and Microsoft/rad-dino for the image encoder.

4.4 Dataset Selection

In this first adaptation, three captions sets were used: MIMIC-CXR NLE captions, raw captions and Regex captions.

The raw captions consist of the ones from the original MIMIC-CXR dataset collected from doctors. An example is seen below.

FINAL REPORT

EXAMINATION: Chest radiograph

INDICATION: ---year-old female with transient hypoxia

TECHNIQUE: Chest PA and lateral

COMPARISON: Portable chest radiograph --- at 03:03

FINDINGS: No significant interval change compared to the prior radiograph performed 1 hr earlier. Endotracheal tube terminates 4.9 cm above the carina. An enteric tube is partially visualized, extending to at least the body of the stomach. Lung volumes are similar. Mild left retrocardiac opacity likely represents atelectasis, although superimposed infection could be considered in the appropriate clinical setting. No new consolidation. Mild pulmonary vascular congestion. No sizable pleural effusion or pneumothorax. Cardiomeastinal contours are normal. No acute osseous abnormalities.

IMPRESSION: No change in probable left retrocardiac opacity and mild pulmonary vascular congestion.

NLE captions ([Kayser et al.](#)) are generated by the NLE model already refereed in Chapter 3. As NLE captions are more general and smaller, they contain less varied vocabulary than the raw captions. So it may overfit quickly. The process to create NLE captions is performed in two steps ([Kayser et al.](#)): Preprocess of raw captions in MIMIC-CXR original dataset to extract relevant information and use a model to transform the cleaner captions to NLE captions. An example is seen below.

Mild left retrocardiac opacity likely represents atelectasis, although superimposed infection could be considered in the appropriate clinical setting.

As the original MIMIC-CXR captions have multiple paragraphs and other information, such as demographic information of patients that is not clinically relevant, they were preprocessed with regex by the methodology outlined in the work ([Erberk Uslu et al.](#)) without passing by NLE model, creating the regex captions dataset. An example of the result is seen below.

No significant interval change compared to the prior radiograph performed 1 hr earlier. Endotracheal tube terminates 4.9 cm above the carina. An enteric tube is partially visualized, extending to at least the body of the stomach. Lung volumes are similar. Mild left retrocardiac opacity likely represents atelectasis, although superimposed infection could be considered in the appropriate clinical setting. No new consolidation. Mild pulmonary vascular congestion. No sizable pleural effusion or pneumothorax. Cardiomeastinal contours are normal. No acute osseous abnormalities. No change in probable left retrocardiac opacity and mild pulmonary vascular congestion.

The parameters of the training process and evaluation of the three sets of captions are summarized in Appendix B. Note that the text decoder is fine-tuned, the mapping network is also trained, while the image encoder is frozen during training. For GPU memory resources issues, a batch size of 16 was used for the original raw captions, while a batch size of 32 was used for the other two sets. Each model was trained for 20 epochs, and the caption generation was performed using beam search with a beam size of 5. An embedding of size 10 was extracted from the image to assist the caption generation process. The model was trained using a cross-entropy loss between the predicted logits and the ground truth captions. The loss overview obtained from the training in each of the three datasets is displayed in Figure 4.11.

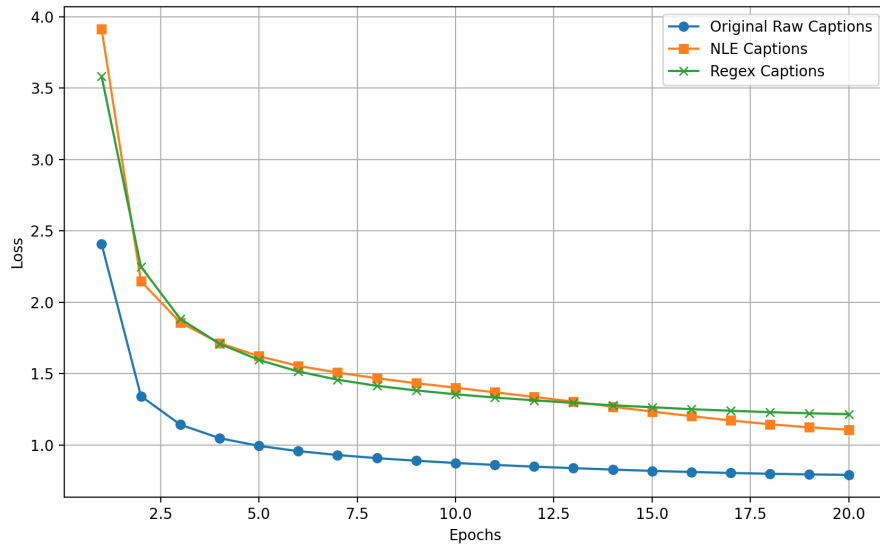


FIGURE 4.11: Dataset Selection Experiments: Epoch Loss Overview.

The model consistently predicts only "FINAL REPORT" across all images for the raw captions, indicating that it is stuck in a local minimum. It minimizes the loss by outputting the report title and whitespace without learning meaningful content. This behaviour demonstrates that the raw captions are unsuitable for training, as further confirmed by the poor evaluation metrics.

Regex and NLE captions have a similar loss convergence over epochs. However, the NLE captions are more repetitive and have fewer clinical details than the Regex processed captions. The metrics to evaluate the performance of each model are depicted in Table 4.5.

TABLE 4.5: Dataset Selection Prediction Metrics in MIMIC-CXR Dataset

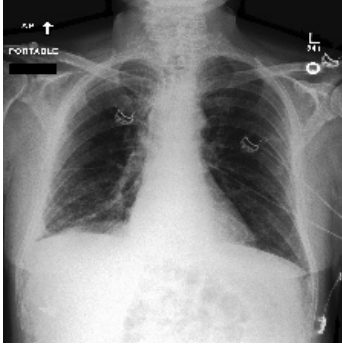
Metric	Raw Reports	NLE Captions	Regex Captions
ROUGE1 Precision	0.020	0.204	0.337
ROUGE1 Recall	1.000	0.221	0.357
ROUGE1 F1	0.039	0.201	0.329
ROUGE2 Precision	0.010	0.059	0.100
ROUGE2 Recall	1.000	0.063	0.108
ROUGE2 F1	0.020	0.058	0.098
ROUGEL Precision	0.020	0.182	0.213
ROUGEL Recall	1.000	0.196	0.226
ROUGEL F1	0.039	0.179	0.207
BERT Precision	0.732	0.872	0.856
BERT Recall	0.899	0.868	0.858
BERT F1	0.807	0.870	0.857
METEOR	0.154	0.141	0.208
CIDEr	0.000	0.217	0.041
BLEU [127]	0.000	0.091	0.156
SembScore [127]	0.040	0.509	0.334
RadGraph Combined [127]	0.000	0.089	0.153
RadCliQ-v0 [127]	5.182	3.517	3.854
RadCliQ-v1 [127]	2.268	1.301	1.431

In most evaluation metrics, the Regex preprocessed captions perform better than those generated with NLE. However, the BERT Score is nearly identical for both methods. While Regex captions perform well in several aspects, they still exhibit a low CIDEr score. This low score indicates that the generated captions overlap less with reference captions, particularly in rare word terms and phrases. Indeed, CIDEr heavily penalizes captions that deviate from the exact style of the ground truth. It als suggests that incorporating a CIDEr score-based reward could be beneficial.

Although some evaluation metrics appear high for the raw captions dataset, these results are misleading. The model overfits by repeatedly generating only two words, presented in the ground truth, which inflates the scores. However, these words lack clinical relevance and do not reflect meaningful caption generation.

Figure 4.12 shows an example of the predictions made by all the experiments for the following ground truth caption:

FINAL REPORT
INDICATION: Status post dilation of esophageal stricture.
COMPARISONS: Chest radiograph Chest radiograph Chest radiograph
FINDINGS: There is no pneumothorax or pneumomediastinum. The cardiomeastinal silhouette is normal. A small right pleural effusion is unchanged. Since the prior radiograph, there has been increased nodular peribronchial opacification, most readily explained by chronic aspiration. Mild hazy opacification at the left base is unchanged and likely represents chronic atelectasis.
IMPRESSION:
 1. No pneumothorax or pneumomediastinum.
 2. Increasing peribronchial opacification at the right base likely represents aspiration, possibly pneumonia.



Regex Caption Prediction: Cardiac silhouette size is normal. The mediastinal and hilar contours are unremarkable. The pulmonary vasculature is normal. Streaky opacities in the lung bases likely reflect areas of atelectasis. No focal consolidation, pleural effusion or pneumothorax is present. There are no acute osseous abnormalities. No acute cardiopulmonary abnormality. No focal consolidation to suggest pneumonia. Mild bibasilar atelectasis. No pleural effusion or pneumothorax.

NLE Caption Prediction: Hazy opacification at the right base is consistent with pleural effusion and elevation of the right hemidiaphragm.

Raw Caption Prediction: FINAL REPORT

FIGURE 4.12: Dataset Selection Predictions for MIMIC-CXR dataset

Captions generated using the NLE dataset are often repetitive and overly generic across multiple images. Although not incorrect, they resemble broad classifications rather than detailed descriptions, likely because the NLE captions were initially generated by a classifier. This problem suggests that the model produces generic outputs that lack alignment with image-specific content and domain-specific language.

The Regex captions dataset is the most promising, as it produces detailed and diverse captions from the images, offering clinical relevance and accuracy.

4.5 Adaptations

This section presents some adaptations of some SOTA models that can be included in the final image captioning model after testing the ability of each adaptation to enhance the model's performance. It will be used the Regex MIMIC-CXR dataset to train and evaluate the models. The main adaptations performed include:

- Incorporate labels to enhance captioning performance while evaluating raw labels and those embedded in the initial prompt. Two labeling tools, Negbio or Chexpert labels, will also be assessed.
- Investigate the benefits of fine-tuning with a CIDEr reward in reinforcement learning to improve captioning accuracy.

- Assess the importance of image features and labels in predicting the generated caption by calculating attention scores.
- Evaluate the benefits of interchanging the order of labels and image embeddings in the last step when generating a caption.
- Analyze the influence of cross-attention instead of simple concatenation between image and label embeddings.
- Investigate the benefits of gathering all image patch embeddings instead of just the CLS token and using cross-attention to extract meaningful embeddings.

Details on the training and prediction settings for each model are provided in Appendix C. Despite minor architectural differences, all models were trained for 20 epochs using MedGPT as the text decoder, Microsoft’s Rad-DINO as the image encoder, and pre-processed captions from the MIMIC-CXR dataset. Caption generation was standardized across models using beam search (beam size 5), a maximum length of 300 tokens, and ‘ / n’ as the stop token, ensuring fair performance comparison.

4.5.1 First Adaptations

An overview of the adaptations performed can be seen in Figure 4.13.

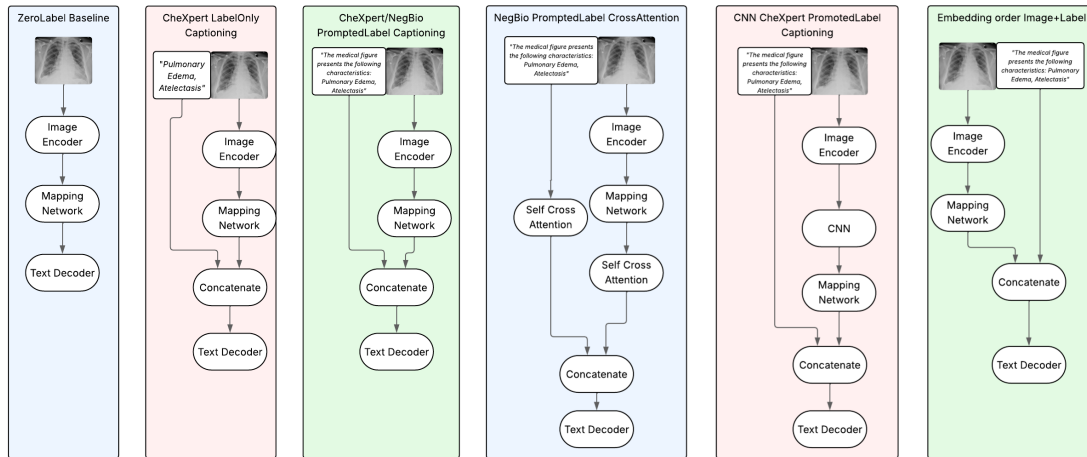


FIGURE 4.13: First Model Adaptations Schema

The base model is depicted in Figure 4.6 which corresponds to the **ZeroLabel Baseline** Experiment.

CheXpert labels, extracted using the labelers described in Kayser et al., were incorporated to improve performance in the **CheXpert LabelOnly Captioning** Experiment.

The labels were embedded, comma-separated, and concatenated with the image embeddings—after projection into the text decoder’s embedding space. This combined input was then passed to the text decoder to generate captions. The corresponding architecture is illustrated in Figure 3.23 in Chapter 3 just by changing the soft prompt and using concatenation instead of cross attention to fuse image embeddings and label embeddings.

Secondly, a soft prompt was used to accommodate the labels *"The medical figure presents the following characteristics:.... "* in **CheXpert Prompted Label Captioning** Experiment. As the prompt contains more tokens, the image embedding prefix length(embedding size after the mapping network) was also increased to size 20 so that the image features can still contribute to generate the caption.

Thirdly, to assess the impact of the label source on training, labels generated by the NegBio model [Kayser et al.](#) were used in place of CheXpert labels in the **NegBio Prompted Label Captioning** Experiment.

Fourthly, it was introduced additional trainable parameters by modifying the fusion strategy of the image embeddings and soft prompt label embeddings in **NegBio Prompted Label CrossAttention** Experiment. Instead of simply concatenating the image embeddings (projected into the text decoder’s feature space) with the soft prompt embeddings, an independent attention layer was applied to each: one for the projected image embeddings and another for the soft prompt embeddings as an intermediate fusion scheme. The outputs of these attention layers were then concatenated. This approach allows a more flexible and dynamic alignment between visual and textual features, enabling the model to capture better the contextual relationships between the image content and the guidance provided by the soft prompt.

Fifthly, the order in which the projected image embeddings and soft prompt with label text embeddings are concatenated is modified in **EmbeddingOrder Image+Label** Experiment. This modification allows for detecting any difference, mainly in the importance given to the labels, so they have more impact on the training process.

Lastly, in the **CNN CheXpert Prompted Label Captioning** Experiment, instead of relying solely on the CLS token, we extracted all patch embeddings from the last hidden state of the image encoder. A Convolutional Neural Network (CNN) then processed these patches to produce a 1024-dimensional embedding per image. This embedding was subsequently projected into the text decoder’s embedding space using the same mapping network. This approach captures more detailed visual information by leveraging

the complete set of patches, enabling a richer and potentially more informative caption generation.

The training loss evolution for all the adaptation experiments described before is displayed in Figure 4.14.

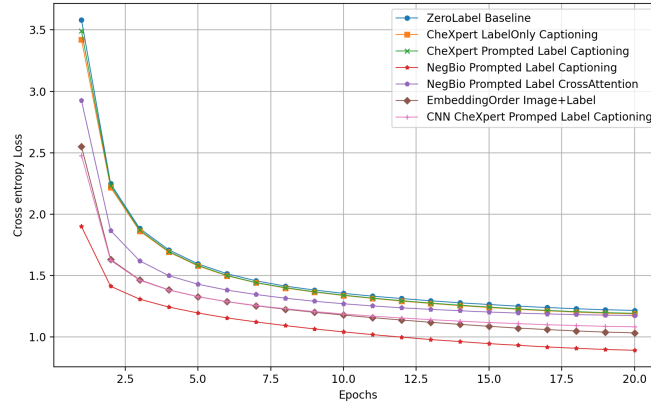


FIGURE 4.14: Adaptations: Epoch Loss Overview

By comparing ZeroLabel Baseline Experiment and CheXpert LabelOnly Captioning Experiment, we conclude that the loss over the first epochs is very similar. However, the advantage of the labels is seen from epoch 10 forward. So, training for more time will benefit the model with labels in terms of loss.

When comparing the CheXpert Prompted Label Captioning Experiment with the CheXpert LabelOnly Captioning Experiment, we see a similar loss behavior over the epochs.

Discussing the ChexPert Prompted Label Captioning Experiment and the NegBio Prompted Label Captioning Experiment, we conclude that the convergence of the model is faster for NegBio labels. However, the metrics are very similar in both experiments as will be seen next.

Comparing CheXpert Prompted Label Captioning and Embedding Order Image+Label Experiments that only differ from the concatenation order of the projected image features and label embeddings, we see that using the image first and then the label embeddings enables a faster convergence.

The CNN CheXpert Prompted Label Captioning loss overview suggests that extracting a large feature embedding from the image helps detect every detail displaying a faster convergence than CheXpert Prompted Label Captioning.

The prediction metrics of the adaptations are displayed in Table 4.6.

By comparing the ZeroLabel Baseline Experiment and the CheXpert LabelOnly Captioning Experiment, we conclude that an improvement in overall performance is observed

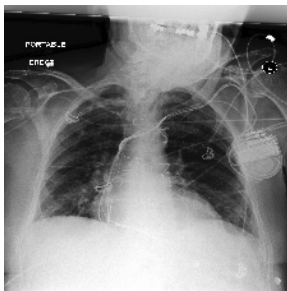
TABLE 4.6: Model Adaptations Predictions metrics in MIMIC-CXR Regex Dataset

Metric	ZeroLabel Baseline	CheXpert LabelOnly Captioning	CheXpert Prompted Label Captioning	NegBio Prompted Label Captioning	NegBio Prompted Label CrossAttention	EmbeddingOrder Image+Label	CNN CheXpert Prompted Label Captioning
ROUGE1 Precision	0.337	0.367	0.362	0.395	0.239	0.335	0.382
ROUGE1 Recall	0.357	0.351	0.360	0.322	0.355	0.335	0.348
ROUGE1 F1	0.329	0.340	0.344	0.337	0.261	0.317	0.347
ROUGE2 Precision	0.101	0.122	0.118	0.127	0.074	0.105	0.127
ROUGE2 Recall	0.108	0.117	0.118	0.104	0.111	0.106	0.116
ROUGE2 F1	0.098	0.113	0.112	0.108	0.081	0.100	0.116
ROUGEL Precision	0.213	0.236	0.234	0.247	0.159	0.212	0.245
ROUGEL Recall	0.226	0.224	0.231	0.199	0.242	0.212	0.221
ROUGEL F1	0.207	0.217	0.221	0.209	0.174	0.200	0.221
BERT Precision	0.856	0.857	0.858	0.859	0.839	0.853	0.860
BERT Recall	0.858	0.857	0.858	0.842	0.832	0.843	0.854
BERT F1	0.857	0.857	0.858	0.850	0.835	0.847	0.857
METEOR	0.208	0.210	0.213	0.192	0.178	0.194	0.209
CIDEr	0.041	0.052	0.072	0.044	0.033	0.053	0.067
BLEU [127]	0.156	0.170	0.173	0.167	0.107	0.156	0.177
SembScore [127]	0.334	0.413	0.417	0.428	0.435	0.423	0.430
RadGraph Combined [127]	0.153	0.176	0.175	0.180	0.159	0.176	0.181
RadCliQ-v0 [127]	3.854	3.684	3.649	3.718	3.899	3.778	3.625
RadCliQ-v1 [127]	1.431	1.326	1.307	1.345	1.476	1.384	1.290

using the labels, that is clear in domain specific metrics. However, the CIDEr score remains too small, suggesting that using Self-Critical Sequence Training with glscider score as a reward might reach higher scores.

On the other hand, CNN CheXpert Prompted Label Captioning outperforms the other models mainly in domain specific metrics, suggesting that using CNN enables to extract more relevant information, which boosts the model performance. Exchanging the order of image embedding, projected features, and label embeddings was not favored in terms of metrics.

The ground truth of the following image is displayed in Figure 4.15, and the predictions of multiple adaptations are depicted in the box below.



Ground Truth

Findings:

Patient is status **post median sternotomy and coronary artery bypass surgery**. ICD remains **in place** as well as a **right PICC**. Cardiac silhouette is **mildly enlarged**, and accompanied by **mild pulmonary vascular congestion**. Persistent **patchy right basilar opacity** and new **patchy left lower lobe opacity** as well as a **persistent linear area of atelectasis in the left lower lobe**. The etiology of the **basilar opacities** is uncertain, but could represent **aspiration, infectious pneumonia, or a dependent distribution of edema** in the setting of known upper lobe predominant **emphysema**.

FIGURE 4.15: MIMI CXR Test Set Example

ZeroLabel Baseline: The patient is status post median sternotomy and CABG. A left-sided pacer device is noted with leads terminating in the right atrium and right ventricle. The heart is moderately enlarged. The mediastinal and hilar contours appear unchanged. There is no pleural effusion or pneumothorax. Streaky bibasilar airspace opacities suggest minor atelectasis. There is no evidence for pulmonary edema. There are no acute osseous abnormalities. No evidence of pulmonary edema. Mild cardiomegaly. Mild bibasilar atelectasis. CheXpert LabelOnly Captioning: In comparison with the study of ____, the monitoring and support devices remain in place. There is some indistinctness of pulmonary vessels consistent with some elevation of pulmonary venous pressure. In the appropriate clinical setting, superimposed pneumonia would have to be seriously considered. However, in the appropriate clinical setting, superimposed pneumonia would have to be seriously considered. Opacification at the right base is consistent with volume loss in the lower lobe and pleural effusion. CheXpert Prompted Label Captioning: In comparison with the study of ____, the monitoring and support devices remain in place. Continued enlargement of the cardiac silhouette with pulmonary edema. Opacification at the right base is consistent with volume loss in the lower lobe and pleural effusion. In the appropriate clinical setting, supervening pneumonia would have to be considered. Less prominent changes are seen at the left base. The monitoring and support devices remain in place. NegBio Prompted Label Captioning: In comparison with the study of ____, the monitoring and support devices remain in place. Continued enlargement of the cardiac silhouette with pulmonary edema. More focal opacification at the right base could reflect developing consolidation in the appropriate clinical setting. In the appropriate clinical setting, superimposed pneumonia would have to be considered. The left hemidiaphragm is not well seen, consistent with volume loss in the lower lobe and possible small effusion. No evidence of pneumothorax. Continued surveillance of the patient after cardiac surgery is recommended. Continued monitoring and support devices are in place. NegBio Prompted Label CrossAttention: Right PICC line tip is at the level of mid SVC. No pneumothorax. Mild cardiomegaly is stable. Mediastinal and hilar contours are stable. Pulmonary vasculature is engorged but there is no overt pulmonary edema. Bibasilar opacities likely reflect atelectasis. No pleural effusion or pneumothorax. Right PICC line tip at the level of mid SVC. Mild cardiomegaly stable. EmbeddingOrder Image+Label:Left-sided AICD, and Transvenous Right Atrial and Right ventricular pacer leads are in standard placements. Patient is status post median sternotomy and CABG. Mild cardiomegaly is accompanied by pulmonary vascular congestion and minimal interstitial edema. New patchy and linear bibasilar opacities have developed in the interval, and may reflect areas of atelectasis, aspiration, and less likely developing pneumonia. Short-term followup radiographs may be helpful in this regard. No large pleural effusion or pneumothorax is present. 1. New patchy and linear bibasilar opacities, which may reflect atelectasis, aspiration, and less likely developing pneumonia. 2. Mild cardiomegaly accompanied by pulmonary vascular congestion and minimal interstitial edema. 3. 4. Transvenous right atrial and right ventricular pacer leads are in standard placements. 5. Mild cardiomegaly. 6. Mild interstitial edema. 7. No pneumothorax. CNN CheXpert Prompted Label Captioning: As compared to the previous radiograph,the patient has received a right PICC line. The course of the line is unremarkable, the tip of the line projects over the mid SVC. The course of the PICC line is unremarkable, the tip of the catheter projects over the mid SVC. The lung volumes remain low.The pre-existing parenchymal opacities at the lung bases have minimally increased in extent and severity.

The metrics in NegBio Prompted Label Captioning and NegBio Prompted Label Cross-Attention Experiments are similar in both models, so there is no significant improvement in using self cross-attention to before combining image embeddings and label embeddings.

The metrics in NegBio Prompted Label Captioning Experiment and EmbeddingOrder Image+Label Experiment are very similar, although when evaluating the generated captions, it was observed that the phrase “as compared with previous study...” no longer appears in EmbeddingOrder Image+Label Experiment, even though it is frequently present in the training set. This appearance of this words suggests that the label soft prompt should be incorporated closely with the generated caption after projecting the image embedding to reduce the redundancy and common phrasing found in the training captions. This approach helps to minimize repetition and encourages the model to generate more diverse and relevant captions.

In fact, EmbeddingOrder Image+Label Experiment along with NegBio Prompted Label Captioning is one of the best caption predictions for the example shown, both in clinical content, with details well aligned with ground truth caption.

CNN CheXpert Prompted Label Captioning Experiment prediction in the example above has focused on the medical tools but lacks some acute clinical details.

Overall, the most promising adaptations extracted from this set of experiments are: using NegBio labels with prompted labels, simple concatenation with image embedding + label embedding in text decoder and using CNN to extract image features with larger dimensions.

4.5.2 Second Adaptations

A new group of adaptations were performed as depicted in Figure 4.16:

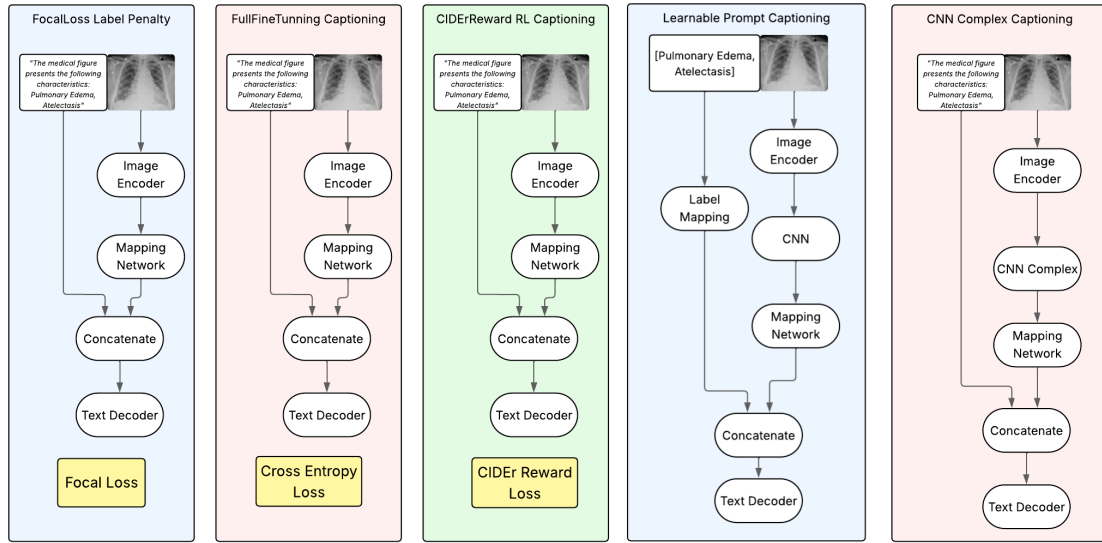


FIGURE 4.16: Second Model Adaptations Schema

Firstly, the model was fine-tuned from the previous NegBio Prompted Label Captioning model by using CIDEr reward in a Reinforcement Learning approach for more 10 epochs. This experiment comprised the **CIDErReward RL Captioning** Experiment.

Then **FullFineTuning Captioning** Experiments consists in training the image encoder, text decoder, and mapping network. This complete training allows us to have a general overview of the training process, which is helpful to recover and analyze gradients of the multiple sub-models, as well as the attention given to the image and label embedding when generating captions over the epochs.

After, a denominated **FocalLoss LabelPenalty** Experiment was performed, employing a focal loss that benefits if the labels are presented in the generated caption. The Focal Loss allows the model to pay more attention to the labels during the caption generation process and increases the influence of the labels on the final model prediction.

Then, a **CNNComplex Captioning** model was trained that uses some of the previous successful adaptations and increases the complexity of CNN backbone used to obtain image features embeddings from all given patches extracted from the image encoder. This was done mainly due to previously observed benefits of using all patch features instead of only the CLS token to extract more relevant information from the image.

Lastly, it was trained a **Learnable Prompt Captioning** model that, instead of using a fixed predefined prompt containing labels, uses a learnable prompt for each set of labels learned during the training process. This increases the model’s freedom to learn the best prompt tailored to each image class, enhancing the captioning performance.

The loss overview for the multiple adaptations are displayed in Figures 4.17, 4.18 and 4.19.

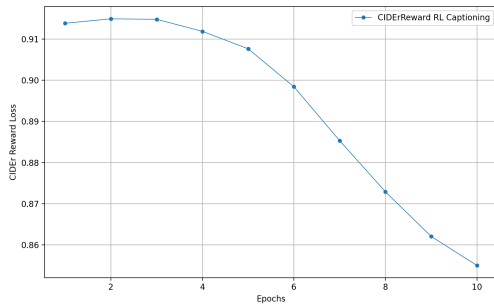


FIGURE 4.17: Epoch Loss Overview - 10 epochs - CIDEr Reward RL Captioning

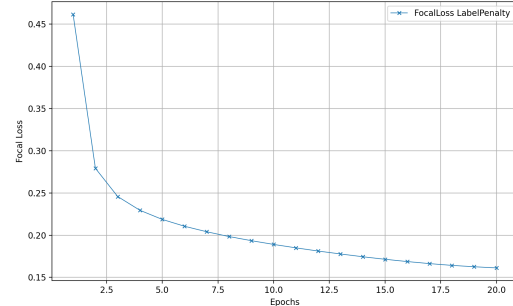


FIGURE 4.18: Epoch Loss Overview - 20 epochs - Focal-Loss LabelPenalty

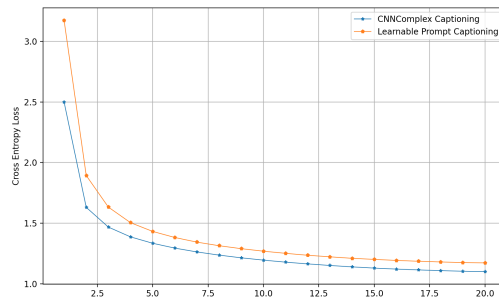


FIGURE 4.19: Epoch Loss Overview - 20 epochs

From the CIDErReward RL Captioning loss overview, it can be observed that the loss initially stagnates but then continues to decrease. This decrease suggests that incorporating the CIDEr score as part of the training process, rather than using it alone, could be a beneficial strategy.

Using CNN Complex Captioning allows a faster convergence and lower loss values using a fixed prompt rather than Learnable Prompt Captioning. Using a learnable prompt solely (without an initialization) leads to too many degrees of freedom, which can make more difficult the learning process.

Using FocalLoss LabelPenalty, we also get a decrease in loss, but it requires a larger number of epochs to achieve convergence.

The prediction metrics obtained for this new set of adaptations are depicted in Figure 4.7.

TABLE 4.7: Model Adaptations Predictions metrics in MIMIC-CXR Regex Dataset

Metric	CIDErReward RL Captioning	FullFineTuning Captioning	FocalLoss LabelPenalty	CNNComplex Captioning	Learnable Prompt Captioning
ROUGE1 Precision	0.225	0.184	0.390	0.385	0.358
ROUGE1 Recall	0.452	0.401	0.310	0.351	0.359
ROUGE1 F1	0.289	0.228	0.329	0.351	0.342
ROUGE2 Precision	0.071	0.064	0.122	0.127	0.112
ROUGE2 Recall	0.145	0.139	0.097	0.115	0.113
ROUGE2 F1	0.092	0.079	0.103	0.116	0.107
ROUGEL Precision	0.152	0.127	0.244	0.245	0.225
ROUGEL Recall	0.305	0.291	0.192	0.222	0.225
ROUGEL F1	0.195	0.158	0.204	0.223	0.214
BERT Precision	0.844	0.830	0.859	0.861	0.860
BERT Recall	0.870	0.859	0.817	0.856	0.857
BERT F1	0.857	0.844	0.837	0.858	0.858
METEOR	0.247	0.193	0.180	0.213	0.214
CIDEr	0.024	0.029	0.033	0.059	0.055
BLEU [127]	0.092	0.090	0.158	0.176	0.168
SembScore [127]	0.327	0.256	0.422	0.442	0.356
RadGraph Combined [127]	0.139	0.124	0.179	0.180	0.169
RadCliQ-v0 [127]	3.881	4.240	3.833	3.597	3.768
RadCliQ-v1 [127]	1.476	1.677	1.412	1.276	1.374

CNNComplex Captioning clearly outperforms the other models in almost all metrics, mainly in domain specific metrics. This indicates that meaningful embeddings extracted from the image are crucial to improve captioning performance.

On the other hand, the CIDErReward RL Captioning approach leads to improvements across all recall-based metrics and METEOR. This improvement indicates that the semantic quality and clinical relevance of the generated captions have improved despite slight declines in some other metrics. Optimizing for CIDEr aligns better with human judgment and retrieval-based evaluations, suggesting that the model generates more meaningful

and contextually appropriate content, even if it sacrifices performance on some surface-level measures.

The ground truth of the following image is shown in Figure 4.20, and below are the multiple model’s caption predictions.

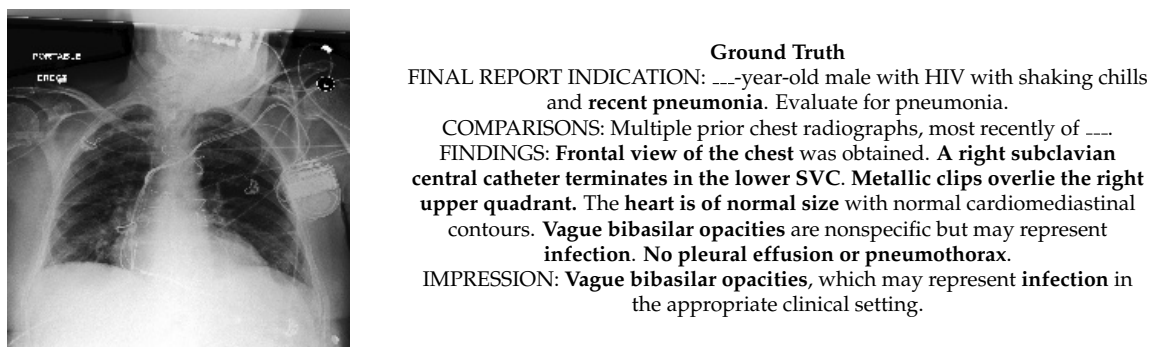


FIGURE 4.20: MIMI CXR Test Set Example

CIDeReward RL Captioning: Right-sided central venous line tip is at the level of mid SVC. Heart size and mediastinum are stable. Right basal opacity most likely represent atelectasis but superimposed infection cannot be excluded in the appropriate clinical setting. No pleural effusion or pneumothorax is seen. FullFineTuning Captioning: There is no pleural effusion or pneumothorax. The cardiomeastinal silhouette is normal. There is no free air under the diaphragm. No free air under the diaphragm FocalLoss LabelPenalty: Right subclavian PICC line terminates in the mid superior vena cava. Cardiomeastinal contours are within normal limits. Minimal patchy opacity in the right lung base likely reflects atelectasis. No focal consolidation, pleural effusion or pneumothorax. Right subclavian PICC line terminates in the mid superior vena cava. These findings were communicated via telephone by Dr. --- to Dr. --- at --- on ---, 10 minutes after discovery. CNNComplex Captioning: In comparison with the study of ---, there is little overall change. Cardiac silhouette is within normal limits and there is no vascular congestion or pleural effusion. Opacification at the left base is consistent with volume loss in the lower lobe and possible small effusion. In the appropriate clinical setting, supervening pneumonia would have to be considered. Right subclavian PICC line extends to the mid portion of the SVC. No evidence of acute focal pneumonia. If the condition of the patient permits, a lateral view would be most helpful. If the condition of the patient permits, a frontal view would be most helpful. Learnable Prompt Captioning: Heart size and mediastinum are stable. Right PICC line tip is at the level of lower SVC. There is no pneumothorax or pleural effusion. There is no pulmonary edema.
--

CIDeReward RL Captioning Experiment provides the most clinically aligned summary — accurate and cautious — though it slightly overstates the diagnosis (atelectasis vs vague opacity). This slight overstate suggests that CIDe optimization encourages the production of focused and content-rich descriptions, potentially by rewarding n-gram overlap with the reference captions, which may naturally favor brevity and clarity over verbosity.

The FocalLoss LabelPenalty experiment excels at capturing detailed structural elements of the image but tends to hallucinate in the metadata or rely heavily on common sentence patterns. In contrast, the CNNComplex Captioning approach generates captions with strong comparative language—characteristic of radiology reports—but is prone to including boilerplate text or incorrect anatomical references, such as “left base.”

Learnable Prompt Captioning Experiment and FullFineTuning Captioning Experiment are concise and readable but miss several clinical details or hedge phrases.

The main benefits extracted from this adaptations is to use a CNN complex model to extract relevant features and maybe using a learnable label prompt although the way embeddings are generated from labels needs to be improved.

4.5.2.1 Gradients Analysis

To perform gradient analysis, the FullFineTuning Captioning Experiment was used to analyze the degree of convergence of multiple models and select the best approach to proceed. The mean normalized value of each model's gradients (in absolute value) was calculated for that and it was computed the relative value in each epoch relative to the first epoch (epoch 0) as shown in Figure 4.21.

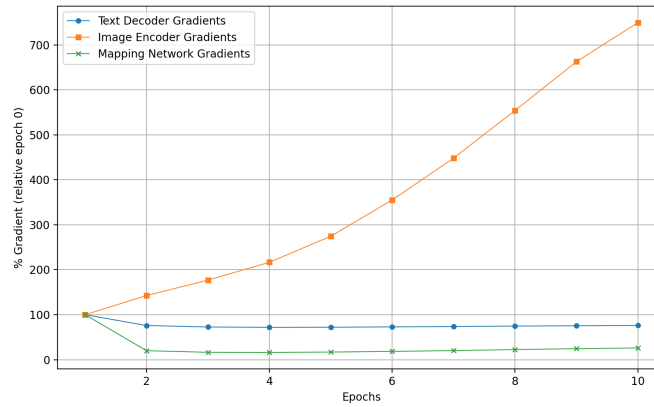


FIGURE 4.21: Full FineTunning Captioning Experiment: Gradient Analysis during training

It is seen that the mapping network quickly converges so that the gradient is closer to zero and stagnates quickly. Although the gradient is higher than the mapping network, the text decoder quickly goes to a constant value. So, it could achieve a state where it cannot learn more. However, the image encoder is the one that, after 10 epochs, the absolute gradient change rate keeps increasing. The curve suggests it will achieve a maximum (negative curvature at the end) and tend to smaller values. So, the image encoder is the submodel component that needs to be trained more and also forces the mapping and text decoder training during the image encoder convergence.

4.5.2.2 Attention Analysis

Using the same model used to analyze the gradients, the attention score was calculated for both the labels and image components in the final embedding array that feeds the text

decoder model. Attention was analyzed in absolute value relative to epoch 0 as shown in Figures 4.22 and 4.23.

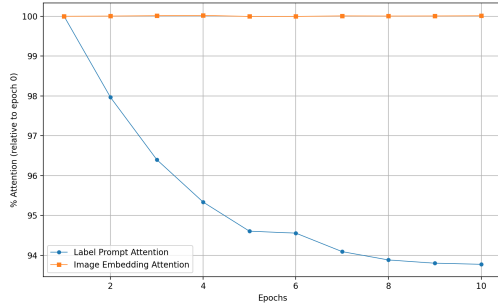


FIGURE 4.22: Attention score (relative epoch 0)

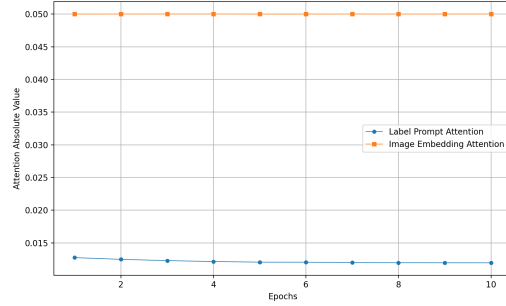


FIGURE 4.23: Attention score (absolute)

It can be seen from the absolute attention score that the image embeddings have approximately five times more attention than the label embedding in the generation of the caption. The relative attention score suggests that the labels are given lower attention over the epochs, while the image attention is almost constant. This fact suggests that the information extracted from labels decreases in importance and relevance over the epochs, mainly because those embeddings do not capture the relevant information as they are supposed to. This problem requires a different approach to extract features from the labels that are clinically relevant and can aid in the caption generation process. To address this issue different label inclusion techniques in the model will be tested in the next section.

4.6 Label Conditioned Image Captioning

The incorporation of labels into the image captioning process can be achieved through three distinct strategies, as described in Section 3: Early Label Conditioned Captioning (Figure 3.24), Intermediate Label Conditioned Captioning (Figure 3.23), and Late Label Conditioned Captioning (Figure 3.25).

As previously mentioned in Section 3, traditional labeling tools such as NegBio or CheXpert cannot be employed directly for training models within the Late Label Conditioned Captioning framework, as they do not support gradient-based backpropagation. To address this limitation, we adopt a transformer-based model called BioBERT, introduced by Lee et al. and fine tune it to the multilabel classification task.

In our approach, the BioBERT was fine-tuned for 30 epochs using ground-truth labels derived from the NegBio labeler. After fine-tuning, it was integrated into the training pipeline of the Late Label Conditioned Captioning model by enabling gradient backpropagation, thereby guiding the caption generation process with label-informed supervision. The results from the BioBERT classifier fine tuning process can be seen in Appendix D.

Each label conditioned model was trained for 20 epochs, with cross-entropy loss and using CNN to extract higher dimension features from the images using the same caption generation methodology and parameters. Appendix D depicts a detailed description of the model parameters.

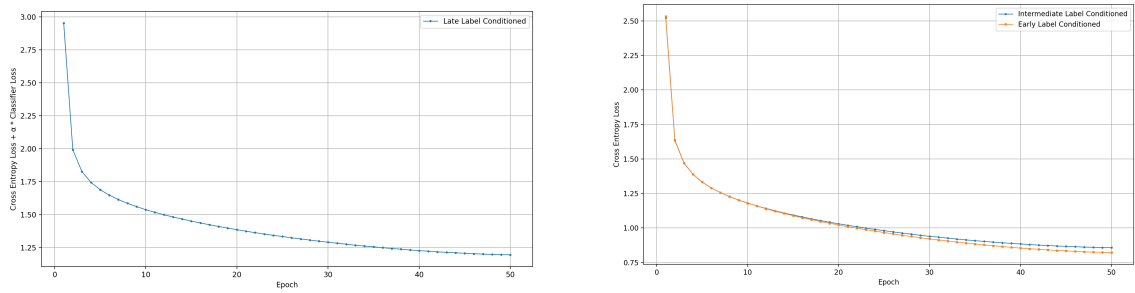


FIGURE 4.24: Label Conditioned Captioning Models: Loss Overview

From the loss overview in Figure 4.24, it is seen that for early and intermediate label-conditioned captioning models the convergence is very similar.

The resulting metrics of each model are shown in Table 4.8. In this final analysis, each label-conditioned model will be compared with the same model without the label condition. It will also compare three baseline models trained in the same MIMIC-CXR Regex dataset, BLIP, ClipCap and MediCap, with the parameters depicted in Appendix A. The evaluation will be done in the MIMIC-CXR dataset, which will be used for training. Zero-shot predictions are also going to be made in the NIH dataset.

BLIP was the model with the worst performance in the MIMIC-CXR Regex dataset. Regarding domain-specific metrics and BLUE score, the label-conditioned captioners outperform the baseline models. Baseline ClipCap model outperforms the label conditional captioning models in Rouge recall metrics, while MediCap performed better in BERT Recall and F1 score, as well as Meteor.

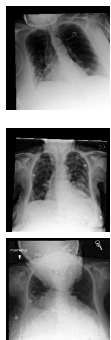
By comparing label fusion approaches, traditional Natural Language Preprocessing (NLP) metrics perform better in the intermediate label conditioned model. Regarding domain-specific metrics, early label conditioned model outperforms all the other ones as

TABLE 4.8: Ablation study: Predictions metrics in MIMIC-CXR Regex Dataset

Metric	Early Label Conditional	Intermediate Label Conditional	Late Label Conditional (BioBERT Classifier)	X-Capfusion without Label	ClipCap (CC)	BLIP	MediCap
ROUGE1 Precision	0.375	0.382	0.386	0.361	0.246	0.184	0.295
ROUGE1 Recall	0.351	0.338	0.291	0.353	0.425	0.114	0.415
ROUGE1 F1	0.345	0.342	0.316	0.341	0.301	0.136	0.331
ROUGE2 Precision	0.127	0.127	0.124	0.116	0.074	0.012	0.094
ROUGE2 Recall	0.117	0.111	0.092	0.113	0.128	0.007	0.131
ROUGE2 F1	0.116	0.115	0.101	0.109	0.090	0.009	0.105
ROUGE1 Precision	0.237	0.242	0.242	0.228	0.162	1.555	0.190
ROUGE1 Recall	0.220	0.213	0.180	0.221	0.278	0.095	0.267
ROUGE1 F1	0.217	0.215	0.196	0.215	0.197	0.114	0.213
BERT Precision	0.860	0.860	0.857	0.858	0.846	0.809	0.855
BERT Recall	0.852	0.845	0.827	0.856	0.867	0.757	0.868
BERT F1	0.855	0.851	0.841	0.857	0.856	0.782	0.861
METEOR	0.213	0.205	0.174	0.210	0.234	0.063	0.240
CIDEr	0.057	0.050	0.056	0.0310	0.025	0.00	0.057
BLEU [127]	0.178	0.160	0.155	0.170	0.107	0.022	0.143
SembScore [127]	0.451	0.394	0.330	0.342	0.267	0.127	0.328
RadGraph Combined [127]	0.181	0.162	0.157	0.165	0.130	0.042	0.146
RadCliQ-v0 [127]	3.590	3.774	4.036	3.810	4.000	5.598	3.810
RadCliQ-v1 [127]	1.270	1.384	1.530	1.397	1.534	2.475	1.414

well as in CIDEr and BLEU scores. In fact early label conditional model has more parameters to train due to the need to predict a labels embedding for each set of labels. This allows a higher performance that is clearly seen in medical domain evaluation metrics.

It is worth showing the decoded label prompt extracted in early label conditioned captioning as in Figure 4.25.



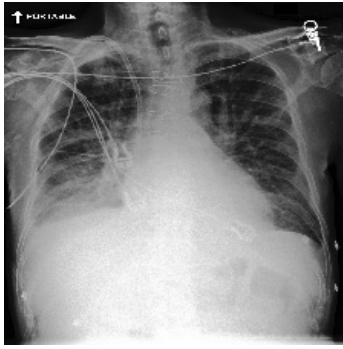
(a) **Decoded Soft Prompt:** 'The leftward and the mediastinal clips are seen in the upper lobesinterstitial lesion of the mediastinal and of the midline. There is increased density in the left hila. The right lung is a very little. It is normal and there is a very subtle increase in the right base. The lung volumes are not significantly changed. The left lung is normal in size. The heart is not as prominent. There is no focal consolidation, but the heart is top normal.'

(b) **Decoded Soft Prompt:** 'The heart size is not as it was there was a prior chest radiograph. The heart is not as it was on the prior. I have both of which is normal. There is a rounded density at the right apical opacification in the left lung apex, not included in the hemidiaphragm. No pleural effusion. The heart is top- the right. The heart is normal in the upper lobe are.'

(c) **Decoded Soft Prompt:** 'As on the right side the aortic and left hemidiaphragmatic hernia. There is a tortuous aorta. This is again seen on the right. This is not well evaluated. There is a new moderate hialal hernia. This is normal. This is a normal finding. The heart size is normal. The cardiomediastinal and mediastinal contours are normal. The heart size is normal. There is a.'

FIGURE 4.25: Decoded Soft Prompt examples: Early Label Conditioned Captioning

In Figure 4.26 it shows an example of an image caption pair in MIMIC-CXR dataset as well as the predicted captions performa by the models.



Ground Truth FINAL REPORT

INDICATION: Status post CABG with dyspnea. Evaluate for edema.

COMPARISONS: Chest radiograph, ____ Chest radiograph, ____

FINDINGS: A right internal jugular hemodialysis catheter ends in the right atrium. The size of the cardiac silhouette is at the upper limits of normal. Sternal wires are intact. A moderate right pleural effusion is slightly bigger. There has been slight increase in the pulmonary edema. Opacification at the right base persists and may be a pneumonia. There is no pneumothorax.

IMPRESSION: 1. Increasing pulmonary edema and enlargement of the moderate right pleural effusion. 2. Possible right lower lobe pneumonia is unchanged.

FIGURE 4.26: MIMIC CXR Test Set Example

X-CapFusion without Label: The patient is status post median sternotomy and CABG. Heart size remains mildly enlarged. Mediastinal and hilar contours are unchanged. Pulmonary vasculature is not engorged. Small bilateral pleural effusions are present, slightly decreased in size compared to the previous exam. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities. Small bilateral pleural effusions, slightly decreased in size compared to the previous exam. Patchy bibasilar airspace opacities likely reflect atelectasis. Infection, however, is not excluded in the correct clinical setting. Mild bibasilar atelectasis. Mild cardiomegaly. No pulmonary edema.

Intermediate Label Conditional: The patient is status post median sternotomy and CABG. Moderate cardiomegaly is re-demonstrated. The mediastinal and hilar contours are unchanged with atherosclerotic calcifications noted at the aortic knob. Mild pulmonary edema is new in the interval with small bilateral pleural effusions. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities. Mild congestive heart failure with small bilateral pleural effusions and mild pulmonary edema. Probable bibasilar atelectasis. No pneumothorax.

Early Label Conditional: The patient is status post median sternotomy and CABG. Moderate cardiomegaly is re-demonstrated. The mediastinal and hilar contours are unchanged. There is mild pulmonary edema with small bilateral pleural effusions. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities. Mild congestive heart failure with small bilateral pleural effusions and mild pulmonary edema. Probable bibasilar atelectasis. Followup radiographs after diuresis are recommended to ensure resolution of this finding.

Late Label Conditional: Right internal jugular central venous catheter tip terminates at the junction of the SVC and right atrium. No pneumothorax is identified. Moderate cardiomegaly is re-demonstrated. The mediastinal and hilar contours are unchanged. Pulmonary vasculature is not engorged. Patchy opacities in the lung bases likely reflect areas of atelectasis. No large pleural effusion or pneumothorax is present. There are no acute osseous abnormalities. Right internal jugular central venous catheter tip at the junction of the SVC and right atrium. No pneumothorax. Bibasilar atelectasis. Unchanged moderate cardiomegaly. No pulmonary edema. Right internal jugular central venous catheter tip terminates at the junction of the SVC and right atrium.

BLIP: patient is status post median sternotomy heart size remains mild to moderately enlarged with a left ventricular predominance the aorta is tortuous.

ClipCap: In comparison with the study of ____, the monitoring and support devices are essentially unchanged. Continued enlargement of the cardiac silhouette with mild elevation of pulmonary venous pressure. Retrocardiac opacification is consistent with volume loss in the left lower lobe and pleural effusion. Indistinctness of engorged pulmonary vessels is consistent

MediCap: In comparison with the study of ____, the monitoring and support devices remain in place. Continued enlargement of the cardiac silhouette with evidence of elevated pulmonary venous pressure. Retrocardiac opacification is consistent with volume loss in the left lower lobe and pleural effusion.

The Early and Intermediate Label Conditional models tend to be safer and more reliable, providing accurate information, although they may omit certain relevant findings. In contrast, the Late Label Conditional model is more detailed but occasionally includes inaccurate information. The models without labels generally produce sparse and often incorrect content. Specifically, BLIP generates very limited information and fails to mention pulmonary findings. Overall, label-guided models (especially Early and Intermediate) strike a better balance between safety and informativeness, while unlabeled or weakly supervised models struggle with completeness and accuracy.

Table 4.9 shows the evaluation metrics on the NIH dataset for the multiple models in the zero-shot scenario.

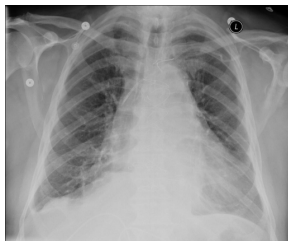
In the zero-shot scenario, MediCap outperforms all the other models, and BLIP performs the worst. However, X-CapFusion without Labels also performed quite well in

TABLE 4.9: Ablation Study: Predictions metrics in NIH Dataset

Metric	Early Label Conditional	Intermediate Label Conditional	Late Label Conditional (BioBERT Classifier)	X-Capfusion without Label	ClipCap (CC)	BLIP	MediCap
ROUGE1 Precision	0.294	0.347	0.429	0.442	0.355	0.137	0.384
ROUGE1 Recall	0.254	0.193	0.207	0.262	0.355	0.047	0.339
ROUGE1 F1	0.252	0.236	0.265	0.309	0.339	0.067	0.342
ROUGE2 Precision	0.087	0.077	0.123	0.133	0.099	0.014	0.115
ROUGE2 Recall	0.078	0.043	0.058	0.077	0.100	0.004	0.104
ROUGE2 F1	0.076	0.053	0.075	0.091	0.095	0.006	0.104
ROUGE1 Precision	0.211	0.225	0.280	0.289	0.244	0.114	0.262
ROUGE1 Recall	0.181	0.122	0.132	0.169	0.239	0.039	0.227
ROUGE1 F1	0.180	0.151	0.170	0.200	0.230	0.056	0.231
BERT Precision	0.849	0.855	0.865	0.867	0.862	0.815	0.868
BERT Recall	0.862	0.837	0.811	0.843	0.861	0.744	0.859
BERT F1	0.855	0.846	0.837	0.854	0.861	0.777	0.864
METEOR	0.140	0.108	0.123	0.155	0.206	0.027	0.209
CIDEr	0.012	0.007	0.005	0.014	0.030	0.000	0.027
BLEU [127]	0.122	0.109	0.155	0.155	0.170	0.012	0.180
SembScore [127]	0.413	0.232	0.330	0.431	0.484	0.180	0.504
RadGraph Combined [127]	0.189	0.108	0.157	0.211	0.183	0.038	0.189
RadCliQ-v0 [127]	3.705	4.294	4.036	3.560	3.386	5.650	3.315
RadCliQ-v1 [127]	1.348	1.705	1.530	1.251	1.162	2.510	1.118

zero shoot scenario mainly in RadGraph Combined. It is worth noting that the caption predictions in the NIH dataset of MediCap tend to be very repetitive and generic, which explains the high metrics, although they lack the details in the images. Label-conditional models also performed well in zero-shot scenarios, mainly the Early Label Conditional Captioning (in domain-specific metrics).

Figure 4.27 displas an example of MIMIC-CXR dataset along with the predictions from the evaluated models.



Ground Truth
Hypoxia, aortic stenosis Stable cardiomegaly. Improved aeration in the left lung base. Right basilar airspace disease is unchanged. No pneumothorax or pleural effusion. Right internal jugular central line has been removed. Left subclavian line is stable.

FIGURE 4.27: NIH Test Set Example

X-Capfusion without Label Frontal and lateral views of the chest were obtained. The patient is status post median sternotomy and CABG. There is blunting of the left costophrenic angle which may be due to a trace pleural effusion. No focal consolidation is seen. There is no evidence of pneumothorax. The cardiac and mediastinal silhouettes are stable. Slight blunting of the left costophrenic angle may be due to a trace pleural effusion. No focal consolidation.

Intermediate Label Conditional: In comparison with the study of ---, there is little overall change. Again there is enlargement of the cardiac silhouette with pulmonary vascular congestion. Opacification at the left base is consistent with volume loss in the lower lobe and pleural effusion. Blunting of the costophrenic angles could reflect small effusions or pleural thickening. No evidence of acute focal pneumonia. The right IJ catheter extends to the mid portion of the SVC. 1. 2. Blunting of the costophrenic angles could reflect small pleural effusions or pleural thickening. 3. 4. 5. This preliminary report was reviewed with Dr. --- radiologist.

Early Label Conditional: The heart is normal in size. There is no pleural effusion or pneumothorax. There is no pulmonary edema.

Late Label Conditional: Patient is status post median sternotomy and CABG. Left-sided Port-A-Cath tip terminates in the proximal right atrium. Heart size remains mildly enlarged. Mediastinal and hilar contours are unchanged. Pulmonary vasculature is not engorged. Small bilateral pleural effusions are demonstrated with bibasilar patchy opacities, likely atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities. Small bilateral pleural effusions with bibasilar patchy opacities, likely atelectasis. Infection, however, is not excluded in the correct clinical setting. Unchanged small bilateral pleural effusions with bibasilar atelectasis.

BLIP: patient is status post median sternotomy heart size remains mild to moderately enlarged with a left ventricular predominance the aorta is tortuous mediastinal

MediCap: Patient is status post median sternotomy and CABG. Heart size is mildly enlarged. Mediastinal and hilar contours are unremarkable. Pulmonary vasculature is not engorged. Minimal streaky opacities in the lung bases likely reflect areas of atelectasis. No focal consolidation, pleural effusion or pneumothorax is present. No acute osseous abnormality is visualized. Mild bibasilar atelectatic changes.

In this example, although the NIH dataset example ground truth caption contains limited information, the evaluation is guided by expert medical interpretation. The key finding—left-sided pleural effusion—is captured by the Intermediate Label Conditional and the model without label, while both the Early and Late Label Conditional models fail to mention it, which is a significant omission. The Intermediate model aligns more closely with the ground truth and appears more reliable overall, even though it introduces some additional hypotheses (e.g., pleural thickening). The Late model, while more detailed, overlooks critical findings. The Early model is conservative but misses important elements. Overall, Intermediate provides the best balance of accuracy and coverage, especially when aligned with annotated findings, while unlabeled and Late models are less consistent in this case.

4.7 Conclusions

The models generally have a small literal matching (low BLEU and CIDEr) and a high semantic matching (high BERT and METEOR). The domain-specific metrics also indicate a good alignment with medical domain semantics.

The results obtained from the baseline models in zero-shot scenarios highlight the importance of domain-specific training and proper dataset size.

Overall, while MediCap demonstrated the highest metrics scores mainly in the zero-shot scenario, it was clear that models trained on general datasets, such as ClipCap and BLIP, struggled with medical image captioning tasks by generating repetitive or nonsensical captions that lacked medical relevance in the zero-shot scenario. In a few shot scenarios, BLIP and ClipCap performed well, while DiCO performed poorly in both frameworks.

The MIMIC-CXR NLE dataset generally enables better model learning compared to ROCO due to its greater size. An observation during analysis revealed that many images in the MIMIC-CXR NLE dataset shared identical labels, suggesting that introducing caption variability through text data augmentation techniques could introduce more variability in training data and enhance model performance.

After performing experiments with multiple text decoders and image encoders using ClipCap architecture and evaluating Grad-CAM, the Rad+MedGPT pair was selected.

Also, three sets of captions were trained, and regex preprocessed captions allowed a better performance, producing detailed and diverse captions from the images and offering clinical relevance and accuracy.

Multiple adaptations to the base model were developed mainly using Negbio labels, complex CNN to extract image features, and ultimately including labels in early, intermediate and late fusion techniques.

An ablation study comparing label-conditioned captioning models with baseline approaches reveals that, in few-shot scenarios, label conditioned models consistently outperform several SOTA image captioning systems. This performance is achieved while maintaining a lightweight architecture and significantly faster training times. Notably, the early label conditioned model demonstrates superior performance, distinguishing itself as the most effective among the variants tested. However, in zero-shot settings, particularly on the NIH dataset, models such as MediCap and X-CapFusion (without label conditioning) outperform label conditioned approaches. This suggests that while label conditioning enhances the model’s ability to capture fine-grained details and subtle clinical nuances, these advantages are most pronounced when some degree of fine-tuning or supervision is available. In truly zero-shot scenarios, the reliance on label-specific conditioning may limit generalization, whereas more general-purpose captioning models exhibit better robustness.

Chapter 5

Conclusions

This chapter presents a summary of the main contributions of this thesis. In addition to that, we will also enumerate some possible future work for subsequent research on medical image captioning that is worth a detailed analysis.

5.1 Contributions

This thesis has allowed a comparison of four state-of-the-art (SOTA) image captioning models of recent research. They were evaluated in both zero-shot and few-shot scenarios in MIMIC-CXR Natural Language Explanation (NLE) and Radiology Objects in COntext (ROCO) datasets. The results have reinforced the necessity to retrain image captioning models on medical datasets to use them in zero-shot scenarios. The most promising models were ClipCap due to pre-trained submodels such as image encoders and text decoders, enabling much faster training, and BLIP, due to the high metrics obtained in a few-shot training. MediCap also performed decently, mainly in zero-shot (it was the only model pre-trained in the medical domain), although BLIP surpassed it in few-shot training.

Due to the goal of using a lightweight model, it was chosen to proceed ClipCap architecture and multiple adaptations were performed to enhance the performance based also on other SOTA models evaluated. Firstly, multiple pre-trained image encoders and text decoders were analyzed to select the most aligned ones. A Grad-CAM analysis was also performed on image encoders, and finally, rad-dino was selected as the image encoder and MedGPT as the text decoder. The reason behind this choice was the good performance in evaluation metrics and the lower cross-entropy loss during training, indicating that they might be more aligned in the medical domain. Although using GPT2 as a text

decoder performed very well, MedGPT model, being trained in the medical corpus, contains more domain-specific language.

Then the different MIMIC-CXR datasets versions were analyzed: NLE captions, Raw captions and Preprocessed regex captions (our contribution). This choice highly influences the model training and final performance. It was found out that raw captions performed very poorly due to the noise in captions (white spaces, titles, demographic information etc.), while NLE captions tend to produce very repetitive words in the dataset and less detailed captions. In fact, the preprocessed regex captions were the best in terms of metrics and in producing clearer, detailed, varied and concise captions.

Finally, multiple adaptations were performed to the base model previously selected. The main improvement of the base model is the inclusion of labels in the intermediate label conditioning caption with a fixed soft prompt. The inclusion of NegBio labels performed slightly better than CheXpert labels. Besides, using a complex Convolutional Neural Network (CNN) to extract higher embedding dimensions from images instead of extracting embeddings just from the CLS token enables a significant increase in performance.

Finally, the three label conditional fusion techniques, early, intermediate, and late label conditional captioning, were analyzed along with an ablation study with the SOTA models. In general, the power of labels and conditional captioning is more clearly seen in a few-shot scenarios, showing a superior performance than some of SOTA models, especially Early Label Conditioned Captioning. In the zero-shot scenario, the labels' conditional captioning superior performance is not seen as equivalent to some of the baseline models. In fact, Label conditioning proves effective in improving the model's sensitivity to detailed and nuanced clinical features. However, its benefits are most evident when the model has access to some form of fine-tuning or supervised adaptation. In contrast, in strict zero-shot settings, label-dependent models may face limitations in generalization. Here, more generic captioning approaches tend to show greater resilience and adaptability across unseen data.

5.2 Future Work

Future work should explore several directions to enhance the training and performance of medical image captioning systems, particularly in the context of domain adaptation, efficient training, and model interpretability. One key avenue involves leveraging LoRA adapters, which allow for efficient fine-tuning of large pretrained models by inserting lightweight, trainable modules into their architecture. This method drastically reduces computational and memory requirements, supporting scalable adaptation across tasks, domains, and even private datasets—critical for medical applications.

Additionally, further experiments are needed to fine-tune training parameters specific to each architecture under investigation. Given the varying behavior and capacity of models such as GPT-2-based decoders (e.g., DeepSeek variants), it is important to systematically optimize hyperparameters—such as learning rates, dropout rates, and regularization techniques—for each configuration. This targeted tuning would help uncover the most effective setups, especially when balancing model simplicity with expressive power.

Another direction is to investigate fine-tuning the image encoder itself. Preliminary gradient analysis suggests that enhancing the encoder could improve the model’s ability to capture subtle, high-value features in medical imagery that are essential for accurate caption generation. In parallel, attention analysis has shown that label-derived information currently plays a smaller role compared to visual embeddings. To address this imbalance, future research should explore methods to strengthen label influence—such as adjusting the weighting factor (α) in the late label conditional captioning loss or employing more sophisticated label integration strategies.

Exploring cross-attention mechanisms at different points within the transformer architecture also represents a promising path. Implementing cross-attention either at the final stages or within intermediate layers may offer improved fusion of visual and textual modalities. Moreover, text conditioning strategies—including tokenizer-level conditioning, embedding injection, or keyform-based controls—should be studied to guide the generation process more explicitly, especially when incorporating structured clinical labels or metadata.

The role of the tokenizer also warrants further analysis. Since tokenization directly influences the model’s input representation, experimenting with domain-specific tokenizers trained on medical corpora could lead to better alignment between text input and visual

features. Data-driven tokenizer learning approaches or tokenizer fine-tuning may further improve performance by adapting vocabulary segmentation to domain-specific language use.

On the representational side, embedding compression is another critical area to explore. The current use of 1024-dimensional embeddings may be inefficient or redundant for some use cases. Alternatives such as compressing embeddings, splitting them into multiple sub-vectors, or tailoring embedding strategies to downstream tasks could yield more efficient and focused representations. Similarly, evaluation methods should be refined using domain-specific metrics, like a medical variant of BERTScore, to ensure that model outputs are assessed based on clinical relevance and language accuracy.

Finally, integrating external knowledge sources such as CLIP features, zero-shot models, or adapters can enrich the model’s understanding, especially in low-data scenarios. Incorporating such external priors—potentially via late fusion strategies—offers a flexible and modular approach to augmenting caption generation without increasing model complexity.

Together, these directions offer a comprehensive roadmap to build more robust, efficient, and clinically meaningful medical image captioning systems.

Appendix A

Baseline Models

The following section presents the parameters used for training the checkpoints used for zero shot and few shot predictions on MIMIC-CXR and Radiology Objects in COntext (ROCO) datasets.

A.1 Zero shot Predictions

To perform zero-shot predictions using ClipCap model the parameters use correspond to the best performance referenced in [69] and are presented in Table A.1. To perform zero shot image captioning predictions using BLIP model the parameters used correspond to the ones in Table A.2. Its worth noting that during pre training the text encoder and decoder share all parameters except self attention layers. To perform zero shot image captioning predictions using MediCap model the parameters used correspond to the ones depicted in Table A.3. To perform zero shot image captioning predictions using DiCO model the parameters used correspond to the one depicted in Table A.4.

TABLE A.1: ClipCap Parameters

Parameter	Description
Image Encoder	ViT
Text Decoder	GPT2
Prefix Length	10
Clip Length	10
Mapping type	Transformer Based
Normalize Prefix	True
Prefix Dimension	512 (ViT)
Datasets	COCO and ConceptualCaptions
Search strategy	Beam Search

TABLE A.2: BLIP Parameters

Parameter	Description
Image Encoder	ViT B16 (ImageNet)
Text Decoder	BERT
Dataset	14M images incl. COCO, VG, CC12M, SBU
Prompts	"a picture of"
Text Generation	Nucleus search (p=0.9)

TABLE A.3: MediCap Parameters

Parameter	Description
Image Encoder	CvT-21
Text Decoder	DistilGPT2
Dataset	extended ROCO
Training	TF + SCST (BERTScore)
Generation	greedy + beam search

TABLE A.4: DiCO Parameters

Parameter	Description
Image Encoder	ViT-L/14
Text Decoder	Transformer (3 layers, 512, 8 heads)
Dataset	COCO Karpathy splits
Training	CLIP-Score + PAC-Score
Generation	Beam search

A.2 Few shot Predictions

To perform few shot predictions using ClipCap model the parameters use correspond to the best performance refereed in [69] and are presented in Table A.5. BLIP model was pretrained in a very large dataset and fine tuned on MIMIC and ROCO for the image captioning task. For the pretrained model it was used the BLIP w/ ViT-B checkpoint with 129 M parameters. To perform few shot image captioning predictions using this model the parameters used correspond to the ones in Table A.7. To perform few shot image captioning predictions using MediCap model the parameters used correspond to the ones depicted in Table A.6. To perform few shot image captioning predictions using DiCO model the parameters used correspond to the one depicted in Table A.8.

TABLE A.5: ClipCap Parameters

Parameter	Description
Image Encoder	ViT-B-32
Text Decoder	GPT2
Prefix Length	10
Clip Length	10
Mapping type	Transformer Based (8 layers)
Normalize Prefix	True
Prefix Dimension	512 for ViT
Datasets	MIMIC CXR and ROCO
Search strategy	Beam Search
Epochs	100
Learning rate	2e-5

TABLE A.6: MediCap Parameters

Parameter	Description
Image Encoder	CvT-21
Text Decoder	DistilGPT2
Dataset	MIMIC and ROCO
Training Methods	Teacher Forcing (TF) and SCST (BERT score reward)
Text Generation	Greedy search and Beam search
Epochs	10 (TF) and 3 (SCST)
Decoder Length	Maximum 256
Number test beams	4
Early stop	True
Learning rate	5e-5 (TF) and 5e-6 (SCST)

TABLE A.7: BLIP Parameters

Parameter	Description
Image Encoder	ViT B16
Text Decoder	BERT
Image Size	384
Dataset	MIMIC and ROCO
Prompts	"a picture of"
Text Generation	Nucleus search (p=0.9)
Weight Decay	0.05
Learning Rate	1e-5
Epochs	5
Vocabulary size	30522
Image embedding size	768
Batch size	32
Generation Length	5 to 20 tokens
Number Beams	1

TABLE A.8: DiCO Parameters

Parameter	Description
Image Encoder	ViT-L/14
Text Decoder	Transformer (3 layers, hidden size 512 and 8 attention heads)
Dataset	MIMIC and ROCO
Training Methods	Without and With Reward (CLIP-Score and PAC-Score)
Text Generation	Beam search
Number steps	10000 (without reward) and 250000 (with reward)
Generation Length	Maximum 60 tokens
Learning Rate	2.5e-4 (without reward) and 1e-6 (with reward)
Optimizer	Adam
Gradient accumulation	Every 8 steps
Number Beams	5
Batch Size	32 (without reward) and 8 (with reward)
Metric for best model	CIDEr (without reward) and ref-PACScore (with reward)

Appendix B

Pretrained Submodels Selection

The following section presents the parameters used for training the models to choose the best image encoder and text decoder to proceed. The main goal was to use ClipCap architecture where image encoder and text decoder are connected by a transformer based mapping network. It was used multiple image encoders and text decoders that were frozen during training and only train the mapping network. This allows to detected what are the best pretrained models adapted to the medical domain..

To perform training based on ClipCap model the parameters use correspond to the best performance referenced in [69] and are presented in table B.1.

TABLE B.1: Model Training Parameters

Parameter	Description
Image Encoder	rad-dino; BiomedCLIP; ViT-B/32
Text Decoder	gpt-2; MedGPT
Prefix Length	10
Clip Length	10
Mapping type	Transformer Based (8 layers)
Normalize Prefix	True
Prefix Dimension	768 for rad-dino; 768 for BiomedCLIP; 512 for ViT-B/32
Datasets	MIMIC CXR and ROCO
Search strategy	Beam Search
Epochs	10
Learning rate	2e-5
Warmup steps	5000
Batch size	32

Appendix C

Dataset Selection and Adaptations

In the following appendix, it is presented the complete tables of the training parameters used for multiple adaptations performed on base image captioning model to reach the final *X-CapFusion* model.

C.0.1 Dataset Selection

The training parameters used in evaluating the performance with all three sets of captions are shown in figure [C.1](#)

TABLE C.1: Training Parameters for Dataset Selection

Parameter	Description
Dataset	MIMIC-CXR
Batch Size	32
Text Generation	Beam search
Number Beams	5
Generation Length	60
Temperature	1.0
Image Embedding Prefix Length	10 tokens
Transformer Mapping	8 layers
Learning Rate	2e-5
Warmup Steps	5000
Image Prefix Length	10
Prefix Length Clip	10
Prefix Dimension	768
Epochs	20

As it will not be using validation dataset it was added to the training dataset to increase the training set. Given that some studies report multiple images corresponding to

the same caption it was randomly used in each epoch an image in folder of that patient ID and report ID. This allows the same caption to be associated with multiples images of the same study, increasing variability.

The prediction step varies a bit depending on the data used. For Natural Language Explanation (NLE) captions it uses '.' as stop token whereas for the other two it uses '/n' token with post processing by removing repetitive sentences. It was used a maximum length of caption of 67 tokens for NLE and 300 tokens for the other two. No prompt was used in the generation process.

Additionally, in this version, the text decoder is fine-tuned while the image encoder is frozen during training. Note that for GPU memory resources we needed to use 16 batch size for original raw captions!

C.0.2 Adaptations

This section presents the training parameters for multiple adaptations performed in tables [C.2](#) and [C.3](#)

TABLE C.2: Training Parameters for Adaptations

Parameter	ZeroLabel Baseline		CheXpert LabelOnly Captioning		CheXpert Prompted Label Captioning		NegBio Prompted Label Captioning		NegBio Prompted Label CrossAttention		EmbeddingOrder Image+Label	CNN CheXpert Prompted Label Captioning
Dataset	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions		
Batch Size	32		32		32		32		32		32	16
Image Encoder	rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (CLS token)	rad-dino (All patches)
Text Decoder	GPTMed		GPTMed		GPTMed		GPTMed		GPTMed		GPTMed	GPTMed
Image Embedding Prefix Length	10		10		20		20		20		20	20
Transformer Mapping	ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)	CNN + ClipCap Mapping (8 layers)
Learning Rate	2e-5		2e-5		2e-5		2e-5		2e-5		2e-5	2e-5
Warmup Steps	5000		5000		5000		5000		5000		5000	5000
Prefix Length Clip	10		10		10		10		10		10	10
Prefix Dimension	768		768		768		768		768		768	1024
Train Image Encoder	False		False		False		False		False		False	False
Train Mapping	True		True		True		True		True		True	True
Train Text Decoder	True		True		True		True		True		True	True
Label and Image Embedding Fusion	Concatenate		Concatenate		Concatenate		Concatenate		Cross Attention		Concatenate	Concatenate
Text Decoder Input Embedding	Image Embedding		Label Embedding + Image Embedding		Label Embedding + Image Embedding		Label Embedding + Image Embedding		—————		Image Embedding + Label Embedding	Label Embedding + Image Embedding
Labels Model	—————		CheXpert		CheXpert		NegBio		CheXpert		CheXpert	CheXpert
Labels Prompt	—————		f'', 'join(label)."		f''The medical figure presents the following characteristics: ', 'join(label)."		f''The medical figure presents the following characteristics: ', 'join(label)."		f''The medical figure presents the following characteristics: ', 'join(label)."		f''The medical figure presents the following characteristics: ', 'join(label)."	f''The medical figure presents the following characteristics: ', 'join(label)."
Loss	Cross Entropy Loss		Cross Entropy Loss		Cross Entropy Loss		Cross Entropy Loss		Cross Entropy Loss		Cross Entropy Loss	Cross Entropy Loss
Epochs	20		20		20		20		20		20	20
Text Generation	Beam search		Beam search		Beam search		Beam search		Beam search		Beam search	Beam search
Number Beams	5		5		5		5		5		5	5
Generation Length	300		300		300		300		300		300	300
Temperature	1.0		1.0		1.0		1.0		1.0		1.0	1.0
Stop Token	/n		/n		/n		/n		/n		/n	/n

TABLE C.3: Training Parameters for Adaptations

Parameter	CIDErReward RL Captioning		FullFineTuning Captioning		FocalLoss LabelPenalty		CNNComplex Captioning		Learnable Prompt CrossAttention	
Dataset	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions	MIMIC-CXR	Prepro- cessed Captions
Batch Size	32		4		16		16		16	
Image Encoder	rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (CLS token)		rad-dino (All patches)		rad-dino (All patches)	
Text Decoder	GPTMed		GPTMed		GPTMed		GPTMed		GPTMed	
Image Embedding Prefix Length	20		20		20		20		20	
Transformer Mapping	ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		ClipCap Mapping (8 layers)		Complex CNN + ClipCap Mapping (8 layers)		Complex CNN + ClipCap Mapping (8 layers)	
Learning Rate	2e-5		2e-5		2e-5		2e-5		2e-5	
Warmup Steps	5000		5000		5000		5000		5000	
Prefix Length Clip	10		10		10		10		10	
Prefix Dimension	768		768		768		2048		2048	
Train Image Encoder	False		True		False		False		False	
Train Mapping	True		True		True		True		True	
Train Text Decoder	True		True		True		True		True	
Label and Image Embedding Fusion	Concatenate		Concatenate		Concatenate		Concatenate		Concatenate	
Text Decoder Input Embedding	Label Embedding + Image Embedding		Label Embedding + Image Embedding		Label Embedding + Image Embedding		Label Embedding + Image Embedding		Label Embedding + Image Embedding	
Labels Model	Negbio		Chexpert		Negbio		Negbio		Negbio	
Labels Prompt	f"The medical figure presents the following characteristics: ', 'join(label)."		f"The medical figure presents the following characteristics: ', 'join(label)."		f"The medical figure presents the following characteristics: ', 'join(label)."		f"The medical figure presents the following characteristics: ', 'join(label)."		None, use Learnable Label Prompt	
Loss	CIDEr Reward Loss		Cross Entropy Loss		Label Focal Loss		Cross Entropy Loss		Cross Entropy Loss	
Epochs	10		10		20		20		20	
Text Generation	Beam search		Beam search		Beam search		Beam search		Beam search	
Number Beams	5		5		5		5		5	
Generation Length	300		300		300		300		300	
Temperature	1.0		1.0		1.0		1.0		1.0	
Stop Token	/n		/n		/n		/n		/n	

Note that the CIDErReward RL Captioning model used was trained using the checkpoint from NegBio Prompted Label Captioning previously trained.

Appendix D

Label Conditioned Captioning

D.0.1 Caption Multilabel Classifier

This section presents the results obtained from fine-tuning the BioBERT-based caption classifier to align with the 14 multi-label classes defined by the NegBio labeler.

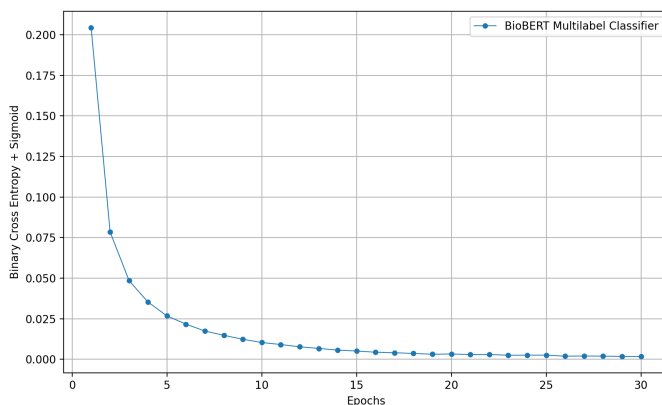


FIGURE D.1: Classifier Epoch Loss Overview

To generate the final predictions, the output probabilities were converted into binary values (0 or 1) using a dynamic thresholding approach. Rather than relying on a fixed threshold or optimizing for accuracy—which is not ideal for imbalanced classification tasks—the threshold was selected based on the value that maximized the $F1 - \beta$ score. This metric offers a more robust evaluation in the presence of class imbalance and better reflects the performance across all labels in the dataset.

The evaluation metrics reported in Table D.1 include micro and macro averages for accuracy, precision, recall, and F1 score, as well as class-specific Area Under the Curve (AUC) values. Micro-averaged metrics aggregate contributions of all classes to provide an overall performance measure, which is particularly useful when classes are imbalanced.

In contrast, macro-averaged metrics compute the metric independently for each class and then average the results, thus treating all classes equally. The AUC values for each class indicate the model’s ability to distinguish between positive and negative instances, with values closer to 1 denoting excellent discrimination performance. Together, these metrics provide a comprehensive assessment of the model’s classification capability on the multi-label MIMIC-CXR dataset.

TABLE D.1: Classifier Evaluation Metrics in MIMIC-CXR Natural Language Explanation (NLE) Dataset.

Metric	Value	Metric	Value
Accuracy micro	0.994	AUC class 7	0.995
Precision micro	0.981	AUC class 8	0.998
Recall micro	0.979	AUC class 9	0.993
F1 micro	0.980	AUC class 10	1.000
Precision macro	0.983	AUC class 11	0.976
Recall macro	0.973	AUC class 12	1.000
F1 macro	0.977	AUC class 13	0.996
AUC class 0	0.999	AUC macro	0.996
AUC class 1	0.999		
AUC class 2	0.998		
AUC class 3	0.999		
AUC class 4	0.998		
AUC class 5	1.000		
AUC class 6	0.991		

Using dynamic thresholds clearly enables a very high performance predicting the multilabels for each caption in MIMIC-CXR dataset.

D.0.2 Label Conditioned Image Captioning

This section presents the training parameters for the label conditioned image captioning models performed in table [D.2](#).

TABLE D.2: Training Parameters for Label Conditioned Models

Parameter	Early Label Conditioned Captioning	Intermediate Label Conditioned Captioning	Late Label Conditioned Captioning (Own Classifier)	Late Label Conditioned Captioning (BERT Classifier)
Dataset	MIMIC-CXR Preprocessed Captions	MIMIC-CXR Preprocessed Captions	MIMIC-CXR Preprocessed Captions	MIMIC-CXR Preprocessed Captions
Batch Size	4	4	8	4
Image Encoder	rad-dino (CLS token)	rad-dino (CLS token)	rad-dino (CLS token)	rad-dino (All patches)
Text Decoder	GPTMed	GPTMed	GPTMed	GPTMed
Image Embedding Prefix Length	10	10	20	20
Transformer Mapping	Complex CNN + ClipCap Mapping (8 layers)	Complex CNN + ClipCap Mapping (8 layers)	Complex CNN + ClipCap Mapping (8 layers)	Complex CNN + ClipCap Mapping (8 layers)
Learning Rate	2e-5	2e-5	2e-5	2e-5
Warmup Steps	5000	5000	5000	5000
Prefix Length Clip	10	10	10	10
Prefix Dimension	2048	2048	2048	2048
Train Image Encoder	False	False	False	False
Train Mapping	True	True	True	True
Train Label Mapping	True	—	—	—
Train Text Decoder	True	True	True	True
Label and Image Embedding Fusion	Cross-Attention	Cross-Attention	—	—
Text Decoder Input Embedding	Label Embedding + Image Embed- ding	Label Embedding + Image Embed- ding	Image Embedding	Image Embedding
Labels Model	Negbio	Negbio	Negbio(Own Classifier)	Negbio (BERT Classifier)
Labels Prompt	None, use Learnable Label Prompt	f"The image show:...."	—	—
Loss	Cross Entropy Loss	Cross Entropy Loss	Cross Entropy Loss + Classifier loss	Cross Entropy Loss + Classifier loss
Epochs	50	50	20	50
Text Generation	Beam search	Beam search	Beam search	Beam search
Number Beams	5	5	5	5
Generation Length	300	300	300	300
Temperature	1.0	1.0	1.0	1.0
Stop Token	/n	/n	/n	/n

Bibliography

- [1] Sarah A Abdu, Ahmed H Yousef, and Ashraf Salem. Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion*, 76:204–226, 2021. [Cited on page 15.]
- [2] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, 114:107856, 2021. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2021.107856>. URL <https://www.sciencedirect.com/science/article/pii/S0031320321000431>. [Cited on page 15.]
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. [Cited on page 16.]
- [4] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. *CoRR*, abs/2011.07191, 2020. URL <https://arxiv.org/abs/2011.07191>. [Cited on page 11.]
- [5] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022. [Cited on page 15.]
- [6] Djamila-Romaissa Beddiar, Mourad Ouassalah, and Tapio Seppänen. Automatic captioning for medical imaging (mic): a rapid review of literature. *Artificial intelligence review*, 56(5):4019–4076, 2023. [Cited on page 15.]

- [7] Jordan J. Bird, Diego R. Faria, Cristiano Premebida, Anikó Ekárt, and George Vogiatis. Look and listen: A multi-modality late fusion approach to scene classification for autonomous machines. *CoRR*, abs/2007.10175, 2020. URL <https://arxiv.org/abs/2007.10175>. [Cited on page 11.]
- [8] Said Boulahia, Abdenour Amamra, Mohamed Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32, 11 2021. doi: 10.1007/s00138-021-01249-8. [Cited on page 9.]
- [9] Abdennour Boulesnane, Boutheina Mokhtari, Oumnia Rana Segueni, and Slimane Segueni. Uterine ultrasound image captioning using deep learning techniques, 2024. URL <https://arxiv.org/abs/2411.14039>. [Cited on page 18.]
- [10] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. [Cited on pages 16 and 17.]
- [11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. [Cited on page 27.]
- [12] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021. [Cited on page 27.]
- [13] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. [Cited on page 25.]
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL <https://arxiv.org/abs/1810.04805>. [Cited on pages xviii, 19, 28, and 29.]
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth

- 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL <https://arxiv.org/abs/2010.11929>. [Cited on pages xviii and 27.]
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [Cited on pages 2 and 27.]
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. [Cited on page 19.]
- [18] Ege Erberk Uslu, Emine Sezer, and Zekeriya Anil Guven. Nlp-powered healthcare insights: A comparative analysis for multi-labeling classification with mimic-cxr dataset. *IEEE Access*, 12:67314–67324, 2024. doi: 10.1109/ACCESS.2024.3400007. [Cited on page 60.]
- [19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Dai, Jonghyun Kim, Alaaeldin El-Nouby, Andrei Barbu, Long Mai, et al. Multi-scale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. [Cited on page 27.]
- [20] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15561-1. [Cited on page 11.]
- [21] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020. [Cited on page 15.]
- [22] Mingliang Gao, Jun Jiang, Guofeng Zou, Vijay John, and Zheng Liu. Rgb-d-based object recognition using multimodal convolutional neural networks: a survey. *IEEE access*, 7:43110–43136, 2019. [Cited on page 15.]

- [23] Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International conference on artificial intelligence and statistics*, pages 1287–1296. PMLR, 2020. [Cited on page 21.]
- [24] Yanfeng Gu, Tianzhu Liu, Guoming Gao, Guangbo Ren, Yi Ma, Jocelyn Chanut, and Xiuping Jia. Multimodal hyperspectral remote sensing: An overview and perspective. *Science China Information Sciences*, 64:1–24, 2021. [Cited on page 16.]
- [25] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2916887. [Cited on pages xvii and 9.]
- [26] Ryuji Hamamoto. Application of artificial intelligence for medical research. *Biomolecules*, 11(1), 2021. ISSN 2218-273X. doi: 10.3390/biom11010090. URL <https://www.mdpi.com/2218-273X/11/1/90>. [Cited on pages xvii and 2.]
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. [Cited on page 25.]
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. [Cited on page 18.]
- [29] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024. [Cited on page 3.]
- [30] Haithem Hermessi, Olfa Mourali, and Ezzeddine Zagrouba. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183:108036, 2021. [Cited on page 16.]
- [31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. [Cited on page 27.]

- [32] Daibing Hou, Zijian Zhao, Yuying Liu, Faliang Chang, and Sanyuan Hu. Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, 9:21236–21250, 2021. doi: 10.1109/ACCESS.2021.3056175. [Cited on pages xvii, 18, and 20.]
- [33] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL <http://arxiv.org/abs/1704.04861>. [Cited on page 18.]
- [34] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. [Cited on page 25.]
- [35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243. [Cited on page 18.]
- [36] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. [Cited on page 25.]
- [37] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed research international*, 2022(1):5164970, 2022. [Cited on page 1.]
- [38] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023. [Cited on pages 16 and 17.]
- [39] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, et al. Overview of imageclef 2018: Challenges, datasets

- and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 309–334. Springer, 2018. [Cited on page 17.]
- [40] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL <https://arxiv.org/abs/1901.07031>. [Cited on page 17.]
- [41] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019. [Cited on page 32.]
- [42] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. URL <https://arxiv.org/abs/1901.07042>. [Cited on pages xxi, 30, and 32.]
- [43] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. [Cited on pages 16 and 17.]
- [44] Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartłomiej Papież, and Thomas Lukasiewicz. Explaining chest x-ray pathologies in natural language, 2022. URL <https://arxiv.org/abs/2207.04343>. [Cited on pages xviii, 31, 32, 60, 64, and 65.]
- [45] Joo-Chang Kim and Kyungyong Chung. Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE access*, 8:104933–104943, 2020. [Cited on page 3.]

- [46] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. Medgpt: Medical concept prediction from clinical narratives, 2021. URL <https://arxiv.org/abs/2107.03134>. [Cited on pages 19 and 55.]
- [47] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James T. Teo, and Richard J. B. Dobson. Medgpt: Medical concept prediction from clinical narratives. *CoRR*, abs/2107.03134, 2021. URL <https://arxiv.org/abs/2107.03134>. [Cited on pages 28 and 29.]
- [48] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. [Cited on pages xvii and 25.]
- [49] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL <http://dx.doi.org/10.1093/bioinformatics/btz682>. [Cited on pages 28 and 74.]
- [50] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL <https://arxiv.org/abs/1910.13461>. [Cited on page 19.]
- [51] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. URL <https://arxiv.org/abs/2306.00890>. [Cited on pages 18 and 20.]
- [52] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. [Cited on page 2.]

- [53] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022. [Cited on page 16.]
- [54] Jingcheng Li, Lina Yao, Binghao Li, and Claude Sammut. Distilled mid-fusion transformer networks for multi-modal human activity recognition, 2023. URL <https://arxiv.org/abs/2305.03810>. [Cited on page 11.]
- [55] Junbing Li, Changqing Zhang, Joey Tianyi Zhou, Huazhu Fu, Shuyin Xia, and Qinghua Hu. Deep-lift: Deep label-specific feature learning for image annotation. *IEEE transactions on Cybernetics*, 52(8):7732–7741, 2021. [Cited on page 21.]
- [56] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL <https://arxiv.org/abs/2201.12086>. [Cited on pages xix, 18, 19, 40, 43, and 51.]
- [57] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning, 2022. URL <https://arxiv.org/abs/2201.04676>. [Cited on page 19.]
- [58] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. URL <http://arxiv.org/abs/1908.03557>. [Cited on page 11.]
- [59] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods, 2022. URL <https://arxiv.org/abs/2203.05227>. [Cited on page 14.]
- [60] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023. URL <https://arxiv.org/abs/2209.03430>. [Cited on pages xvii, 7, and 8.]
- [61] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual

- language models, 2024. URL <https://arxiv.org/abs/2312.07533>. [Cited on pages 18 and 19.]
- [62] Chang Liu, Yuanhe Tian, and Yan Song. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*, 2023. [Cited on page 15.]
- [63] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL <https://arxiv.org/abs/2103.14030>. [Cited on page 19.]
- [64] Ze Liu, Yutong Lin, Yuqi Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. [Cited on page 27.]
- [65] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016. [Cited on page 1.]
- [66] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, May 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.106791. URL <http://dx.doi.org/10.1016/j.compbiomed.2023.106791>. [Cited on page 19.]
- [67] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40, 2022. [Cited on pages 15, 18, 21, and 22.]
- [68] Hairo Ulises Miranda-Belmonte and Adrian Pastor Lopez-Monroy. Early fusion of traditional and deep features for irony detection in twitter. In *IberLEF@SEPLN*, 2019. URL <https://api.semanticscholar.org/CorpusID:199448234>. [Cited on page 11.]

- [69] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021. URL <https://arxiv.org/abs/2111.09734>. [Cited on pages [xix](#), [11](#), [12](#), [14](#), [18](#), [19](#), [39](#), [43](#), [51](#), [55](#), [57](#), [85](#), [86](#), and [89](#).]
- [70] Nicholas Moratelli, Davide Caffagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting image captioning training paradigm via direct clip-based optimization, 2024. URL <https://arxiv.org/abs/2408.14547>. [Cited on pages [xix](#), [42](#), and [51](#).]
- [71] Recurrent neural network. Recurrent neural network. https://en.wikipedia.org/wiki/Recurrent_neural_network, 2024. Accessed: 2024-09-04. [Cited on pages [xviii](#) and [27](#).]
- [72] Aaron Nicolson, Jason Dowling, and Bevan Koopman. A concise model for medical image captioning. In *CLEF (Working Notes)*, pages 1611–1619, 2023. [Cited on pages [xix](#), [41](#), and [51](#).]
- [73] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation, 2024. URL <https://arxiv.org/abs/2307.09758>. [Cited on pages [11](#), [12](#), and [18](#).]
- [74] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning*, pages 3809–3818. PMLR, 2018. [Cited on page [21](#).]
- [75] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021. doi: 10.1109/TNNLS.2020.2979670. [Cited on page [14](#).]
- [76] T. Pang, P. Li, and L. Zhao. A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22:48, 2023. doi: 10.1186/s12938-023-01113-y. [Cited on page [15](#).]
- [77] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language*, pages 26–36, 2019. [Cited on pages [15](#), [16](#), [17](#), and [19](#).]

- [78] John Pavlopoulos, Vasiliki Kougia, Ion Androutsopoulos, and Dimitris Papamichail. Diagnostic captioning: a survey. *Knowledge and Information Systems*, 64(7):1691–1722, 2022. [Cited on page 15.]
- [79] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In Danail Stoyanov, Zeike Taylor, Simone Balocco, Raphael Sznitman, Anne Martel, Lena Maier-Hein, Luc Duong, Guillaume Zahnd, Stefanie Demirci, Shadi Albarqouni, Su-Lin Lee, Stefano Moriconi, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, Eric Granger, and Pierre Jannin, editors, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01364-6. [Cited on pages xviii, 33, and 34.]
- [80] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018. [Cited on pages 16 and 17.]
- [81] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, 2017. URL <https://arxiv.org/abs/1712.05898>. [Cited on page 32.]
- [82] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1): 119–130, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL <http://dx.doi.org/10.1038/s42256-024-00965-w>. [Cited on page 55.]

- [83] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL <https://api.semanticscholar.org/CorpusID:49313245>. [Cited on page 19.]
- [84] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. [Cited on pages 28 and 29.]
- [85] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. [Cited on page 55.]
- [86] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL <https://arxiv.org/abs/2103.00020>. [Cited on pages 11, 12, 14, and 19.]
- [87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>. [Cited on pages xviii, 29, 30, and 55.]
- [88] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL <http://arxiv.org/abs/1910.10683>. [Cited on pages 19 and 28.]
- [89] Vishwanatha M Rao, Michael Hla, Michael Moor, Subathra Adithan, Stephen Kwak, Eric J Topol, and Pranav Rajpurkar. Multimodal generative ai for medical image interpretation. *Nature*, 639(8056):888–896, 2025. [Cited on page 2.]
- [90] Gabriel Reale-Nosei, Elvira Amador-Domínguez, and Emilio Serrano. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, page 103264, 2024. [Cited on pages 2, 15, 18, and 22.]

- [91] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning, 2017. URL <https://arxiv.org/abs/1612.00563>. [Cited on pages xx and 46.]
- [92] Richard J Roberts. Pubmed central: The genbank of the published literature, 2001. [Cited on page 33.]
- [93] Roboflow. What is a convolutional neural network?, 2024. URL <https://blog.roboflow.com/what-is-a-convolutional-neural-network/>. Accessed: 2024-09-04. [Cited on pages xviii and 25.]
- [94] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024. [Cited on page 17.]
- [95] Alexander Selivanov, Oleg Y Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V Dylov. Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1):4171, 2023. [Cited on page 18.]
- [96] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL <http://arxiv.org/abs/1610.02391>. [Cited on page 59.]
- [97] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. [Cited on page 21.]
- [98] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023. [Cited on page 16.]
- [99] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506, 2016. doi: 10.1109/CVPR.2016.274. [Cited on page 18.]

- [100] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [Cited on page 25.]
- [101] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL <https://arxiv.org/abs/1409.1556>. [Cited on page 18.]
- [102] Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. Medpix 2.0: A comprehensive multimodal biomedical dataset for advanced ai applications. *arXiv preprint arXiv:2407.02994*, 2024. [Cited on pages 16 and 17.]
- [103] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medcat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020. [Cited on pages 16 and 17.]
- [104] Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. Open domain question answering using early fusion of knowledge bases and text. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1455. URL <https://aclanthology.org/D18-1455>. [Cited on page 11.]
- [105] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL <http://arxiv.org/abs/1409.3215>. [Cited on page 28.]
- [106] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. [Cited on page 25.]

- [107] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1109/CVPR.2016.308. [Cited on page 18.]
- [108] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693, 2020. [Cited on page 22.]
- [109] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL <http://arxiv.org/abs/1905.11946>. [Cited on page 18.]
- [110] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. [Cited on page 25.]
- [111] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL <https://arxiv.org/abs/2405.09818>. [Cited on page 11.]
- [112] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL <https://arxiv.org/abs/2312.11805>. [Cited on page 7.]
- [113] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. [Cited on pages xviii, 2, 14, 26, and 28.]
- [114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL <http://arxiv.org/abs/1706.03762>. [Cited on page 19.]
- [115] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL <https://arxiv.org/abs/1411.5726>. [Cited on page 38.]
- [116] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. *CoRR*, abs/1604.00790, 2016. URL <http://arxiv.org/abs/1604.00790>. [Cited on page 18.]

- [117] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. URL <https://arxiv.org/abs/2205.14100>. [Cited on page 11.]
- [118] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. [Cited on page 2.]
- [119] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL <http://arxiv.org/abs/1609.08144>. [Cited on page 28.]
- [120] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. Cares: A comprehensive benchmark of trustworthiness in medical vision language models, 2024. URL <https://arxiv.org/abs/2406.06007>. [Cited on page 18.]
- [121] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models, 2024. URL <https://arxiv.org/abs/2410.13085>. [Cited on pages xvii, 18, and 21.]
- [122] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models, 2024. URL <https://arxiv.org/abs/2407.05131>. [Cited on page 18.]
- [123] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale

- multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. [Cited on pages 16 and 17.]
- [124] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL <http://arxiv.org/abs/1502.03044>. [Cited on page 18.]
- [125] Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546:126287, 2023. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2023.126287>. URL <https://www.sciencedirect.com/science/article/pii/S0925231223004101>. [Cited on page 15.]
- [126] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020. URL <https://arxiv.org/abs/1906.08237>. [Cited on page 19.]
- [127] Fiona Yu, Masaki Endo, Ramesh Krishnan, Irene Pan, Andy Tsai, Eduardo P. Reis, Edson KUN Fonseca, Hyo Min Hwang Lee, Zeshan S. Hameed Abad, Andrew Y. Ng, Curtis P. Langlotz, Vijay K. Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, Aug 2023. doi: 10.1016/j.patter.2023.100802. [Cited on pages xix, 22, 35, 36, 51, 54, 57, 62, 67, 71, 76, and 78.]
- [128] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. [Cited on page 1.]
- [129] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Rob Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv*, January 2024. URL <https://www.microsoft.com/en-us/research/publication/biomedclip-a-multimodal-biomedical-foundation->

[model-pretrained-from-fifteen-million-scientific-image-text-pairs/](#).

[Cited on page 55.]

- [130] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>. [Cited on pages xix, 37, and 38.]
- [131] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multi-modal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021. [Cited on page 15.]
- [132] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022. URL <https://arxiv.org/abs/2010.00747>. [Cited on page 14.]
- [133] Zhenyu Zhang, Benlu Wang, Weijie Liang, Yizhi Li, Xuechen Guo, Guanhong Wang, Shiyang Li, and Gaoang Wang. Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning, 2023. URL <https://arxiv.org/abs/2311.01004>. [Cited on pages 18 and 20.]
- [134] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 2024. [Cited on page 15.]
- [135] Yulai Zhao, Masatoshi Uehara, Gabriele Scalia, Tommaso Biancalani, Sergey Levine, and Ehsan Hajiramezanali. Adding conditional control to diffusion models with reinforcement learning, 2024. URL <https://arxiv.org/abs/2406.12120>. [Cited on page 14.]
- [136] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019. URL <http://arxiv.org/abs/1909.11059>. [Cited on pages xvii, 18, and 19.]
- [137] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multi-modal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023. [Cited on page 15.]