U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

PhD

PhD

PhD

PhD

3rd CYCLE

FCUP
2026

U.PORTO

Weakly Supervised Multimodal Explanations for Medical Image Classification

Isabel Rio-Torto

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Weakly Supervised Multimodal Explanations for Medical Image Classification

Isabel Rio-Torto
Doctoral Program in Computer Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2026

Weakly Supervised Multimodal Explanations for Medical Image Classification

Isabel Rio-Torto
Thesis carried out as part of the Doctoral Program in Computer Science
Department of Computer Science
2026

Supervisor
Luís F. Teixeira, Associate Professor, Faculty of Engineering of the University of Porto

Co-supervisor
Jaime S. Cardoso, Full Professor, Faculty of Engineering of the University of Porto





Acknowledgements

As I now sit here after this long and tumultuous but, in the end, very enriching, fulfilling, and happy journey, I am certain that there are not enough words to express my deepest gratitude to everyone who, in many different ways, contributed to the successful conclusion of this chapter of my life. Nevertheless, I'll give it my best effort.

I want to start by thanking my supervisors, for helping me navigate this adventure. Prof. Jaime Cardoso will always be a role model of scientific integrity, work ethic, and genuine passion for research. Thank you for your guidance and contagious optimism. Prof. Luís Teixeira, it has been an absolute honour and pleasure to have been your student for the past 12 years (yes, it's already been 12 years, I can hardly believe it as well!). I will be forever grateful for everything you have taught me, for your support, your infinite patience, your empathy, and your unwavering faith in my capabilities. Much more than a mentor, you have become a cherished friend.

I also want to express my gratitude to Prof. Zeynep Akata and the Explainable Machine Learning group at the Tuebingen AI Center for having welcomed me for a research stay. It was a very enriching experience and I want to thank Stephan Alaniz, my *de facto* supervisor during my stay. I have learned immensely from you and it was a true pleasure working with you.

To all my colleagues and friends at the VCMI research group, thank you for encouragement and all that you have taught me. A special thank you to Tiago, Tomé, Pedro, Sara, Wilson, and Margarida for putting up with me and my peculiar brand of craziness.

To all my friends, especially the ones that have known me for more than 25 years (Mariana, Nuno, Sara, and Daniel), I want to express my heartfelt gratitude for always being there, even when I sometimes had to disappear for a while to work. Your friendship means the world to me.

Finally, I want to thank the three most important people in my life, my parents, Graça and Pedro, and my partner in crime, Nuno. This PhD would not have been possible without you. You have been my rocks, my safe havens, throughout this entire journey. I will be forever grateful for your unconditional support and love. It is impossible to put into words how much you mean to me, but I will continue to try everyday to put it into actions.

Isabel Rio-Torto

Resumo

Apesar do sucesso de modelos de aprendizagem profunda na análise de imagens médicas, a sua falta de interpretabilidade continua a ser uma barreira crítica à adoção na prática clínica. Embora já muitos métodos de inteligência artificial explicável tenham sido propostos para mitigar este problema, a maioria explora explicações visuais *post-hoc*, uma vez que são fáceis de aplicar a arquiteturas já existentes e não requerem que os modelos sejam retreinados. Contudo, estes métodos apresentam limitações críticas: identificam quais as características relevantes, mas não explicam porquê, são frequentemente difíceis de interpretar, e a sua fidelidade ao modelo é incerta. Além disso, diferentes utilizadores (por exemplo, médicos vs. pacientes) têm necessidades e preferências distintas, pelo que os métodos de explicabilidade devem ser capazes de se adaptar e gerar diferentes modalidades de explicações. Embora os métodos de inteligência artificial explicável melhorem a interpretabilidade, muitas vezes aumentam o já elevado custo de anotação, especialmente no caso de explicações baseadas em conceitos ou explicações em linguagem natural. Mais importante ainda, para esta última modalidade, depender inteiramente da supervisão de anotações feitas por humanos pode até inviabilizar a fidelidade das explicações.

Esta tese aborda estes desafios através de dois objetivos principais: (i) desenvolver alternativas às explicações visuais *post-hoc* para a classificação de imagens médicas, (ii) sem aumentar o custo de anotação, recorrendo a aprendizagem fracamente supervisionada. Para tal, exploram-se múltiplas modalidades de explicação e paradigmas de aprendizagem.

Em primeiro lugar, focamo-nos nas explicações visuais, investigando o efeito dos mecanismos de atenção no desempenho e na interpretabilidade visual *post-model* das redes neuronais profundas. De seguida, analisamos uma arquitetura intrinsecamente interpretável, as redes B-cos, apresentando o primeiro estudo sistemático da sua adequação a classificação de imagem médica. Embora os mapas B-cos ofereçam interpretabilidade localizada, os resultados revelam limitações importantes relativamente à fidelidade e ao desempenho nas tarefas de classificação.

Em segundo lugar, é proposto um novo paradigma para explicações baseadas em conceitos através do modelo *CBVLM*, que transforma Modelos Visão–Linguagem de Grande Escala em modelos baseados em conceitos. Operando sob supervisão fraca (Aprendizagem em Contexto), o modelo *CBVLM* reduz significativamente os custos de anotação, ao mesmo tempo que supera abordagens de referência tradicionais em diversos conjuntos de dados médicos.

Em terceiro lugar, esta tese contribui para a investigação sobre geração de explicações em linguagem natural ao introduzir o modelo *WeNLEX*, um modelo fracamente supervisionado capaz

de gerar explicações textuais fiéis e adaptáveis ao público-alvo. É de notar que o modelo *WeN-LEX* melhora tanto a explicabilidade como o desempenho preditivo quando integrado no treino do classificador subjacente.

Finalmente, esta tese apresenta o modelo *DeViL*, um modelo capaz de gerar explicações visuais e textuais, através do treino com pares imagem–relatório médico, produzindo mapas de saliência para qualquer palavra ou frase e descrições em linguagem natural das características do modelo.

Em conjunto, estas contribuições demonstram que inteligência artificial médica explicável, fiel e adaptável, pode ser alcançada sem *pipelines* de anotação dispendiosos, através da utilização estratégica de supervisão fraca e explicações multimodais.

Palavras-chave: Inteligência Artificial Explicável, Explicações Multimodais, Explicações Visuais, Explicações em Linguagem Natural, Explicações Visuais e Textuais, Aprendizagem Fracamente Supervisionada, Aprendizagem em Contexto, Modelos Visão–Linguagem de Grande Escala, Classificação de Imagens Médicas

Abstract

Despite the remarkable success of Deep Learning (DL) models in medical image analysis, their lack of interpretability remains a critical barrier to clinical adoption. While many Explainable AI (XAI) methods have been proposed to mitigate this problem, most explore *post-hoc* visual explanations because they are easy to apply to existing architectures and do not require retraining the underlying models. However, these have critical limitations: they highlight which features are relevant but not why, they are often hard to interpret, and their faithfulness to the model is uncertain. Moreover, different users (e.g., clinicians vs. patients) have varying needs and preferences, so explainability methods need to be able to adapt and generate different modalities of explanations. Although XAI methods improve interpretability, they often add to the already high annotation burden, especially for concept-based explanations or Natural Language Explanations (NLEs). More importantly, for the latter, relying too heavily on full human supervision can even render explanations unfaithful.

This thesis addresses these challenges by pursuing two main objectives: (i) to develop alternatives to *post-hoc* visual explanations for medical image classification, (ii) without increasing the annotation burden, leveraging weakly supervised learning. To this end, we explore multiple explanation modalities and training paradigms.

First, we focus on visual explanations, investigating the effect of attention mechanisms on the performance and post-model visual interpretability of Deep Neural Networks (DNNs). We then look into an inherently interpretable architecture, B-cos networks, presenting the first systematic study of their suitability for medical image classification. While B-cos maps offer localised interpretability, results reveal important limitations regarding faithfulness and task-specific performance.

Second, a new paradigm for concept-based explanations is proposed through *CBVLM*, which transforms Large Vision-Language Models (LVLMs) into Concept Bottleneck Models (CBMs). Operating under weak supervision (In-Context Learning (ICL)), *CBVLM* significantly reduces annotation costs while outperforming traditional baselines across diverse medical datasets.

Third, this thesis advances research on the generation of NLEs by introducing *WeNLEX*, a weakly supervised framework capable of generating faithful and audience-adapted textual explanations. Notably, *WeNLEX* enhances both explainability and predictive accuracy when integrated into the training of the underlying classifier.

Finally, this thesis presents *DeViL*, a vision–language explanation framework that aligns visual and textual modalities by training on image–medical report pairs, producing open-vocabulary saliency maps and natural language descriptions of model features.

Collectively, these contributions demonstrate that faithful and adaptable medical XAI can be achieved without costly annotation pipelines, through the strategic use of weak supervision and multimodal explanations.

Keywords: Explainable Artificial Intelligence, Multimodal Explanations, Visual Explanations, Natural Language Explanations, Vision-Language Explanations, Weakly Supervised Learning, In-Context Learning, Large Vision-Language Models, Medical Image Classification

Contents

1 Introduction **1**

1.1 Context and Motivation 1

1.2 Objectives, Contributions, and Document Structure 2

1.3 Publications 3

1.4 Funding 4

2 Visual Explanations **5**

2.1 Attention-based Networks 6

2.1.1 Context and Motivation 6

2.1.2 Background 7

2.1.3 Attention Mechanisms for Computer Vision 8

2.1.4 Attention Mechanisms for Medical Image Classification 11

2.1.5 Case Studies 11

2.1.6 Conclusions and Future Challenges 21

2.2 B-cos Networks 23

2.2.1 Context and Motivation 23

2.2.2 Methodology 23

2.2.3 Results and Discussion 26

2.2.4 Summary and Conclusions 30

3 Concept-based Explanations **31**

3.1 Context and Motivation 31

3.2 Related Work 32

3.2.1 Large Vision-Language Models in the Medical Domain 32

3.2.2 Concept Bottleneck Models 33

3.3 Methodology 34

3.3.1 Concept Detection 34

3.3.2 Disease Diagnosis 35

3.3.3 Few-shot Prompting 36

3.3.4 Demonstration Example Retrieval 36

3.3.5 Answer Extraction 36

3.4 Experimental Setup . . . 37
3.4.1 Datasets . . . 37
3.4.2 LVLMs . . . 38
3.4.3 Baselines . . . 38
3.4.4 Retrieval Module . . . 39
3.4.5 Evaluation Metrics . . . 39
3.5 Results and Analysis . . . 39
3.5.1 Concept Detection Performance . . . 40
3.5.2 Disease Diagnosis Performance . . . 43
3.5.3 Ablation Study: Strategies for ICL Example Selection . . . 45
3.5.4 General Discussion . . . 46
3.5.5 Limitations . . . 47
3.6 Summary and Conclusions . . . 49

4 Natural Language Explanations **50**
4.1 Context and Motivation . . . 51
4.2 Related Work . . . 53
4.3 Dataset . . . 54
4.4 Problem Formulation . . . 55
4.5 Fully Supervised Approach . . . 56
4.5.1 Baselines . . . 56
4.5.2 Proposed Approach . . . 56
4.5.3 Experimental Setup . . . 59
4.5.4 Results and Discussion . . . 60
4.5.5 Summary and Conclusions . . . 63
4.6 Weakly Supervised Approach . . . 63
4.6.1 Proposed Model and Loss Functions . . . 63
4.6.2 Experimental Setup . . . 67
4.6.3 Results . . . 71
4.6.4 Ablations . . . 72
4.6.5 Audience-Adaptable NLEs . . . 73
4.6.6 Summary and Conclusions . . . 74

5 Vision-Language Explanations **76**
5.1 Context and Motivation . . . 77
5.2 Related Work . . . 78
5.2.1 Inherently Interpretable Models . . . 78
5.2.2 *Post-hoc* Saliency . . . 78
5.2.3 Natural Language Explanations . . . 79
5.2.4 Captioning Models . . . 79
5.3 *DeViL*: Decoding Vision features into Language . . . 79

5.3.1 Model Architecture . . . 79
5.3.2 Training Process . . . 80
5.3.3 Dropout . . . 81
5.3.4 Open-vocabulary Saliency . . . 82
5.3.5 Experiments and Results . . . 82
5.3.6 Summary and Conclusions . . . 90
5.4 Understanding Chest X-ray Vision Representations with *DeViL* . . . 91
5.4.1 Methodology . . . 91
5.4.2 Experiments and Results . . . 91
5.4.3 Summary and Conclusions . . . 98

6 Summary and Conclusions **99**

References **102**

A Concept-based Explanations **121**
A.1 Comparison Between Example Selection Strategies - RICES and MMICES . . . 121
A.2 Additional visualisations . . . 122
A.3 Implementation Details . . . 123
A.4 Clinical Concepts . . . 124
A.5 Prompts . . . 124

B Vision-Language Explanations **133**
B.1 Implementation Details . . . 133
B.2 Limitations and Failure Cases . . . 134
B.3 Feature Descriptions at Different Locations . . . 134
B.4 Saliency Maps Across Vision Backbones . . . 134
B.5 Detailed Comparison with MILAN . . . 135

List of Figures

2.1 Block diagram of an attention-based Encoder-Decoder network 8
2.2 Block diagram of the SE attention block 9
2.3 Block diagram of the CBAM 10
2.4 Block diagram of the channel attention module of the CBAM 10
2.5 Block diagram of the spatial attention module of the CBAM 10
2.6 Accuracy on the APTOS2019 dataset, with different percentages of training data 16
2.7 Accuracy on the ISIC2020 dataset, with different percentages of training data . . 16
2.8 Accuracy on the MIMIC-CXR dataset, with different percentages of training data 16
2.9 Examples of LRP and DeepLIFT maps for the APTOS2019 dataset 18
2.10 Examples of LRP and DeepLIFT maps for the ISIC2020 dataset 19
2.11 Examples of LRP and DeepLIFT maps for the MIMIC-CXR dataset 20
2.12 Comparison of explanations given by Input×Gradient, Grad-CAM, and B-cos on original and B-cos networks 29

3.1 Comparison between the proposed methodology (*CBVLM*) and existing methods 33
3.2 Overview of *CBVLM* 35
3.3 Concept detection performance of LVLMs across different *n*-shot settings 41
3.4 Concept detection results per dataset averaged over all, generic, and medical LVLMs 41
3.5 Per-concept F1-scores across datasets averaged over all, generic, and medical LVLMs, in the 4-shot scenario 42
3.6 Influence of the example set size on the concept prediction performance of *CBVLM* 43
3.7 Disease diagnosis performance of LVLMs across different *n*-shot settings 44
3.8 Disease diagnosis results per dataset averaged over all, generic, and medical LVLMs 45
3.9 Concept detection results with 1-shot and several example selection strategies . . 47

4.1 Difference between image captions and explanations 52
4.2 Block diagram of the fully supervised NLE generation architecture 57
4.3 Different PEFT architectures of the NLE generator 58
4.4 Illustrative examples of NLEs generated by all tested models 62
4.5 Architecture of *WeNLEX* 64
4.6 Depiction of the deletion faithfulness metric 69

4.7 Correct CLEV computation 70
4.8 Qualitative examples of *WeNLEX* 75

5.1 *DeViL* generates textual descriptions for feature maps of any arbitrary vision backbone 78
5.2 Architecture diagram of the proposed model, *DeViL* 80
5.3 Qualitative results on MILANNOTATIONS 85
5.4 Captions generated for each image by the CLIP-ResNet50 model 86
5.5 Descriptions generated for CLIP-ViT at different layers 87
5.6 Examples of open-vocabulary saliency maps showcasing that semantics are encoded in different layers 88
5.7 Frequency of most similar patch size for a given *DeViL* description and human caption 88
5.8 Open-vocabulary saliency maps 90
5.9 Qualitative examples of open-vocabulary saliency maps generated by the three *DeViL* versions 97

A.1 Disease diagnosis results with 4-shots and two example selection strategies . . . 122
A.2 Comparison of LVLM concept detection performance across datasets under different n-shot settings 123
A.3 Comparison of LVLM disease diagnosis performance across datasets under different n-shot settings 124
A.4 Comparison of per-concept F1-scores across datasets for each evaluated LVLM . 125
A.5 Example concept detection prompt for the Derm7pt dataset with a multiple-choice concept 126
A.6 Example concept detection prompt for the Derm7pt dataset with a binary concept 128
A.7 Example concept detection prompt for the SkinCon, CORDA, and DDR datasets 128
A.8 Example disease diagnosis prompt for the Derm7pt dataset 129
A.9 Example disease diagnosis prompt for the SkinCon dataset 130
A.10 Example disease diagnosis prompt for the CORDA dataset 131
A.11 Example disease diagnosis prompt for the DDR dataset 132

B.1 Failure cases of *DeViL* for CLIP-RN50 135
B.2 Additional examples of generated localised descriptions for CLIP-ResNet50 on COCO 136
B.3 Additional examples of generated localised descriptions CLIP-ViT on COCO . . 136
B.4 Additional examples of generated localised descriptions for CLIP-ResNet50 on ImageNet 137
B.5 Additional open-vocabulary saliency examples for different vision backbones . . 137
B.6 Additional qualitative examples on the MILANNOTATIONS dataset 138

List of Tables

1.1 Overview of this thesis contributions 2

2.1 Data splits and label meaning for each dataset 13
2.2 Data augmentation parameters used in all experiments 14
2.3 Set of hyperparameters used in all experiments 15
2.4 Parameter count of all tested models 17
2.5 Results on model performance and XAI quality assessment per model across several datasets 28

3.1 Detailed statistics for the datasets used 38
3.2 Comparison of the best performing *CBVLM* with state-of-the-art supervised models and CBM for disease diagnosis and concept detection 48

4.1 Results obtained on the MIMIC-NLE test set 61
4.2 Comparison of *WeNLEX* with the fully supervised baseline 72
4.3 Ablation on the layer used for the feature reconstruction loss 73
4.4 Ablation on the number of NLEs per diagnosis present in the GT NLE database . 73
4.5 Comparison of *WeNLEX* trained on clinician vs. simplified layperson NLEs . . . 74

5.1 Image captioning results on CC3M 84
5.2 Ablation study of the proposed dropout mechanism 84
5.3 Evaluation of MILAN, ClipCap and *DeViL* on MILANNOTATIONS 85
5.4 Per-layer statistics of adjectives, verbs, and number of unique words 88
5.5 Comparison between the ResNet50 model and Grad-CAM for saliency maps generated for different layers 89
5.6 Performance on the SRRG-Impression dataset 93
5.7 Assessment of *DeViL* vs. Grad-CAM saliency map quality on Faithfulness and Localisation metrics 93
5.8 Assessment of *DeViL* vs. NIB saliency map quality on Faithfulness and Localisation Metrics for the XRaySigLIP vision encoder 94
5.9 Open-vocabulary visual saliency evaluation on the MS-CXR dataset 96
5.10 Region-to-text generation results of *DeViL* XRaySigLIP on the MS-CXR dataset 98

A.1 HuggingFace checkpoints of the LVLMs used 126
A.2 List of annotated clinical concepts of each dataset 127

B.1 Results on generalization experiments with the MILANNOTATIONS dataset . . 138

List of Acronyms

AE	Autoencoder
AI	Artificial Intelligence
APTOS	Asia Pacific Tele-Ophthalmology Society
AUC	Area Under the Curve
BACC	Balanced Accuracy
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CAM	Class Activation Map
CBAM	Convolutional Block Attention Module
CBM	Concept Bottleneck Model
CLAT	Concept-based Lesion Aware Transformer
CLEV	Clinical Evidence
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CORDA	COVID Radiographic Images Dataset for AI
COVID	Coronavirus Disease
CT	Computed Tomography
CV	Computer Vision
CXBS	CheXbert Score
CXR	Chest X-ray

DeepLIFT Deep Learning Important Features

DeiT Data-Efficient Image Transformer

DeViL Decoding Vision features into Language

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DNN Deep Neural Network

DPT DenseNet-121 + GPT-2

DR Diabetic Retinopathy

GAN Generative Adversarial Network

GAP Global Average Pooling

GPT Generative Pre-trained Transformer

GPU Graphics Processing Unit

GT Ground Truth

HIPAA Health Insurance Portability and Accountability Act

ICL In-Context Learning

ISIC International Skin Imaging Collaboration

KG Knowledge Graph

LLM Large Language Model

LoRA Low-Rank Adaptation

LRP Layerwise Relevance Propagation

LSTM Long Short-Term Memory

LVLM Large Vision-Language Model

MBE Model Being Explained

METEOR Metric for Evaluation of Translation with Explicit word Ordering

MIL Multiple Instance Learning

MIMIC Medical Information Mart for Intensive Care

MLP Multilayer Perceptron

MMD Maximum Mean Discrepancy

MMICES Mixed Modality In-Context Example Selection

MSE Mean Squared Error

NIB Narrowing Information Bottleneck

NLE Natural Language Explanation

NLG Natural Language Generation

NLP Natural Language Processing

NMT Neural Machine Translation

OPT Open Pre-trained Transformers

PEFT Parameter-Efficient Fine-Tuning

RAG Retrieval-Augmented Generation

RATCHET Radiological Text Captioning for Human Examined Thoraces

ReLU Rectified Linear Unit

RICES Retrieval-based In-Context Example Selection

RKHS Reproducing Kernel Hilbert Space

RL Reinforcement Learning

RNN Recurrent Neural Network

ROUGE Recall-Oriented Understudy for Gisting Evaluation

SE Squeeze-and-Excitation

SigLIP Sigmoid Loss for Language-Image Pretraining

SPICE Semantic Propositional Image Caption Evaluation

SRRG Structured Radiology Report Generation

ViT Vision Transformer

VL Vision-Language

VLM Vision-Language Model

VQA Visual Question Answering

WGAN Wasserstein GAN

WGAN-GP Wasserstein GAN with Gradient Penalty

WSI Whole Slide Image

XAI Explainable AI

Chapter 1

Introduction

1.1 Context and Motivation

The success of DL models, which have reached or even surpassed human performance in several applications, has led to their widespread usage in many areas of our daily lives, and the medical domain is no exception. Despite their potential for the improvement of clinical workflows (e.g., faster and cheaper screening), the lack of interpretability of these models (also called black-box models) has prevented their deployment in real settings.

Research on XAI has emerged to tackle this problem, especially in domains like medicine, given the life-or-death nature of the decisions involved [1]. Generally, XAI methods can be divided into three categories, according to when they are applied relative to the training process of the Model Being Explained (MBE): i) *pre-model*, i.e., before training (e.g., exploratory data analysis and visualisation); ii) *in-model or inherently interpretable models*, i.e., during training (e.g., design of architectures with domain constraints or regularization techniques [2]–[4]); and iii) *post-model/post-hoc*, i.e., after training (e.g., gradient/saliency/perturbation/attention-based methods that quantify the relevance of the features learned by the MBE [5], [6]). Although some authors argue that only inherently interpretable models might guarantee trust and transparency in high-stakes decision fields [7], the majority of works has been focused on *post-hoc* visual explanations, given that they are plug-and-play/model-agnostic and training-free. However, they have concerning limitations [7]–[9]: i) they highlight the relevant features but do not explain why they are relevant in the first place; ii) they are not easily understandable (especially in medical use cases if, e.g., the target audience for the explanation are the patients and not clinicians); and iii) their faithfulness to the model's decision process is not guaranteed.

Furthermore, different users with distinct backgrounds and levels of expertise (e.g., a clinician vs. a patient) have diverse needs and preferences (e.g., some clinicians might prefer looking at the original exams, while others favour reading the exams' textual reports), so systems should be versatile enough to generate and leverage the advantages and complementarity of various modalities of explanations.

Although XAI methods enhance the interpretability of the underlying MBE, they usually increase the already high annotation burden (i.e., datasets need to have clinician-annotated explanations besides the task labels), especially in explanation modalities like concepts or NLEs. In the latter, this is particularly worrisome, since generating these explanations with full supervision of human annotations renders them unfaithful [10].

1.2 Objectives, Contributions, and Document Structure

In light of the challenges identified in the previous section, the goal of this thesis is twofold: to study and develop alternatives to *post-hoc* visual explanations for medical image classification, and to do so without incurring additional annotation costs (via different types of weakly supervised learning). Although some other imaging modalities and classification tasks are explored, every work presented in this thesis considers the multilabel classification of thoracic diseases via Chest X-ray (CXR) analysis as the main use case.

Table 1.1 presents an overview of the contributions made, taking into account the proposed models, the explanation modalities used, the types of weak supervision employed, the tested datasets, and the venues (international journals and conferences) where the research was published.

Table 1.1: Overview of this thesis contributions, considering the proposed models, the explanation modalities explored, the types of (weak) supervision employed, the datasets used, and the venues where the research was published.

Chapter	Proposed Model	Explanation		Type of (Weak) Supervision	Datasets			Publications
		Method	Modality		Name	Classif. Task	Image Modality	
2.1	-	post-model	saliency map	class label	APTOS2019 [11] ISIC2020 [12] MIMIC-CXR [13]	diabetic retinopathy skin lesions pleural effusion	eye fundus dermoscopy CXR	Journal (IEEE Access) [6]
2.2	-	in-model	saliency map	class label	APTOS2019 [11] Herlev [15] ISIC2018 [16], [17] VinDr-CXR [18]	diabetic retinopathy abnormal cells skin lesions thoracic diseases	eye fundus cytology dermoscopy CXR	Conf. (ISBI) [14]
3	CBVLM	post-model	concepts	ICL	CORDA [19] Derm7pt [21] DDR [22] SkinCon [23]	COVID-19 melanoma diabetic retinopathy skin lesion	CXR dermoscopy eye fundus dermoscopy	Journal (CiBM) [20]
4	WeNLEX	in-model or post-model	NLEs	class label and small set of NLEs	MIMIC-NLE [24]	thoracic diseases	CXR	Conf. (IbPRIA) [25] Conf. (MIDL) [26] Journal (submitted) [27]
5	*DeViL*	post-model	image description and saliency maps	image-caption pairs	CC3M [28] COCO [30] ImageNet [31] Places365 [32] MILANNOTATIONS [33]	- - visual objects scenes -	natural images neuron activations	Conf. (GCPR) [29]
				image-medical report pairs	SRRG [34] CheXlocalize [36] MS-CXR [37]	thoracic diseases	CXR	Conf. (CBMS) [35]

Before delving into the document structure, it is worth noting that while some authors distinguish between the terms "interpretability" and "explainability", we use the two terms interchangeably.

We start by focusing on visual explanations (Chapter 2), by first investigating if and how *post-hoc* saliency maps differ in DNNs with and without attention mechanisms (Section 2.1).

Here we first consider explicitly including attention mechanisms into non-attention-based models and then experiment with Vision Transformers (ViTs), which use attention mechanisms as their main building blocks. Afterwards, we study a type of inherently interpretable model called B-cos networks [4] and its applicability to medical image classification (Section 2.2). These novel networks intrinsically produce visual explanations by aligning network weights with inputs during training. Keeping in line with one of the main goals of this thesis, in both works we do not incur in additional annotation costs: the models are trained only with the classification labels.

In Chapter 3 we move on to a different type of explanation modality, concept-based explanations. In this part of the work we transform LVLMs into CBMs, i.e., models that first predict a set of concepts ("bottleneck") and then ground the final classification decision on these predicted concepts. Using LVLMs allows us to exploit their ICL capabilities, thus the approach we propose, *CBVLM*, also works under a weakly supervised paradigm and only requires few examples annotated with Ground Truth (GT) explanations.

Chapter 4 is dedicated to research on NLEs. We start by tackling the problem in the fully supervised scenario, to first demonstrate that it is viable to supervise NLE generation at the sentence level at the output of the Encoder in an Encoder-Decoder framework (Section 4.5). This first step is vital to ensure that NLE generation is still possible when few human-annotated NLEs are available and, therefore, traditional teacher-forcing-based training is not feasible. After empirically showcasing the viability of this approach, we turn to our end-goal: the generation of faithful and plausible NLEs in a weakly supervised manner, through the proposal of a novel model named *WeNLEX* (Section 4.6). *WeNLEX* requires only a few labelled samples to ensure plausibility of the generated NLEs and ensures explanation faithfulness through the reconstruction of the original image features from the NLE embedding space.

Before concluding, Chapter 5 explores vision-language explanations through the proposal of the *DeViL* model. *DeViL* is able to generate aligned vision explanations and textual descriptions, by training on image-caption pairs (i.e., weak labels w.r.t. the task of explanation generation). Besides the original work on natural images, we also test *DeViL* on CXRs by training it on CXR-structured radiology report pairs, and then use *DeViL* to compare different types of CXR encoders.

Finally, Chapter 6 provides the concluding remarks of this thesis, highlighting the contributions made in each chapter, as well as possible future research directions.

1.3 Publications

The contributions of the doctoral research on XAI described in this thesis have been disseminated as part of eight scientific publications. These are (clustered by type and in reverse chronological order):

Articles in Journals

1. Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "WeNLEX: Weakly Supervised Natural Language Explanations for Multilabel Chest X-ray Classification" under revision at IEEE Transactions on Medical Imaging (TMI) [27]

2. Isabel Rio-Torto*, Cristiano Patrício*, Jaime S. Cardoso, Luís F. Teixeira, and João C. Neves, "CBVLM: Training-free Explainable Concept-based Large Vision Language Models for Medical Image Classification" Computers in Biology and Medicine, vol. 198, p. 111 145, 2025 [20]

3. Tiago Gonçalves, Isabel Rio-Torto, Luís F. Teixeira, and Jaime S. Cardoso, "A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?" IEEE Access, vol. 10, pp. 98909-98935, 2022 [6]

Articles in International Conference Proceedings

1. Isabel Rio-Torto, Bruno Coelho, Jaime S. Cardoso, and Luís F. Teixeira, "Understanding Chest X-ray Vision Representations with *DeViL*" accepted at the IEEE International Symposium on Computer-Based Medical Systems (CBMS) [35]

2. Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "Parameter-Efficient Generation of Natural Language Explanations for Chest X-ray Classification" Medical Imaging with Deep Learning (MIDL), vol. 250, pp. 1267–1281, 2024 [26]

3. Isabel Rio-Torto*, Tiago Gonçalves*, Jaime S. Cardoso, and Luís F. Teixeira, "On the Suitability of B-cos Networks for the Medical Domain" Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1-5, 2024 [14]

4. Isabel Rio-Torto*, Meghal Dani*, Stephan Alaniz, and Zeynep Akata, "*DeViL*: Decoding Vision features into Language" Proceedings of the DAGM German Conference on Pattern Recognition (GCPR), pp. 363-377, 2023 [29]

5. Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "From Captions to Explanations: A Multimodal Transformer-based Architecture for Natural Language Explanation Generation" Pattern Recognition and Image Analysis, pp. 54–65, 2022 [25]

* Equal contribution

1.4 Funding

This work has received funding from the Portuguese Foundation for Science and Technology (FCT) through the PhD Grant 2020.07034.BD.

Chapter 2

Visual Explanations

Foreword on Author Contributions

The research work described in this chapter was conducted in collaboration with Tiago Gonçalves, under the supervision of Professors Jaime S. Cardoso, and Luís F. Teixeira.

In the first work [6], the author of this thesis contributed to the implementation of all experiments using the Transformer model, the conceptualization, implementation, and running of the experiments with different percentages of training data, results analysis and discussion, and preparation of the scientific publication.

In the second work [14], the author of this thesis contributed equally with Tiago Gonçalves, as both authors worked on every step of the paper: conceptualization, implementation, results analysis and discussion, and preparation of the scientific publication.

The results of these works have been disseminated in two articles, one in an international journal, and the other in an international conference:

[6] Tiago Gonçalves, Isabel Rio-Torto, Luís F. Teixeira, and Jaime S. Cardoso, "A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?" IEEE Access, vol. 10, pp. 98909-98935, 2022

Code: https://github.com/TiagoFilipeSousaGoncalves/survey-attention-medical-imaging

[14] Isabel Rio-Torto*, Tiago Gonçalves*, Jaime S. Cardoso, and Luís F. Teixeira, "On the Suitability of B-cos Networks for the Medical Domain" Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1-5, 2024

* Equal contribution

Code: https://github.com/icrto/medical-bcos

This chapter focuses on one of the most common explanation modalities in medical image classification: visual explanations (also known as saliency maps). This type of maps show how important each pixel (of the input image) is for the classification decision.

We first explore different types of attention mechanisms, their impact on model predictive performance and complexity, and how they influence *post-hoc* saliency maps. In the second part of the chapter we investigate if B-cos networks [4], a type of inherently interpretable neural network that generates visual explanations by aligning model weights with network inputs, are suitable for the medical domain.

In both sections of the chapter, no direct supervision is used to generate the visual explanations. In the first part we use gradient-based post-model XAI methods, which are model-agnostic and training-free. In the second part, only classification labels are used to train the models and, consequently, produce visual explanations.

2.1 Attention-based Networks

2.1.1 Context and Motivation

The concept of attention has long been studied across psychology, and neuroscience [38]. Its integration into Artificial Intelligence (AI) was largely inspired by biological vision systems, which appear to analyse complex scenes through sequential processing strategies [39]. Early studies connected attention to the problem of visual search, such as identifying a small object within a cluttered environment. Although high-resolution images with a wide field of view could address this challenge, they produce extremely high-dimensional inputs [40]. Consequently, effective mechanisms were required to prioritize relevant regions and features while efficiently allocating computational resources. This led to the development of selective visual attention models in Computer Vision (CV), focusing on identifying regions of interest and relevant feature dimensions [41]. Over time, multiple research directions contributed to the development of modern attention mechanisms in DL [38]–[41]. More recently, advances in computational power and the widespread availability of large-scale datasets have revitalised DL, enabling the integration of learnable attention mechanisms in end-to-end frameworks with significant impact across diverse applications [42]–[45].

In 2017, Vaswani *et al.* [46] proposed the Transformer, a DNN architecture composed solely of attention operations that discard recurrence and convolutions entirely. In the original paper, the authors applied this architecture to the task of Neural Machine Translation (NMT) and explored the use of attention operations as a means of reducing computational complexity when compared to recurrent or convolution operations. The establishment of a similar architecture for CV occurred almost naturally, when Dosovitskiy *et al.* [47] proposed the ViT architecture.

Given the multitude of works using attention mechanisms, we investigate how attention mechanisms are being used for medical applications (including Transformer-based architectures), and how they impact performance, model complexity and interpretability. We approach this topic with a critical view and ask the following research questions:

1. Do attention mechanisms automatically improve the predictive power of DL algorithms for medical image applications?

2. What is the impact of integrating attention mechanisms on model complexity?

3. Can we improve the degree of interpretability of DL models solely through attention mechanisms?

4. How practical is it to design and build attention mechanisms for DL applications?

To answer these, we develop an experimental protocol with three case studies on medical image classification, using established DL architectures with and without state-of-the-art attention blocks and a Transformer-based model. We address the open challenges of interpretability and transparency with the help of *post-hoc* explanation methods and perform a visual analysis of the impact of different attention mechanisms in these outputs.

Thus, the main contributions of this work are: an experimental case study on medical image classification with three different use cases that focus on the integration of attention mechanisms into established DL architectures, the analysis of their predictive power, and a visual assessment of their saliency maps generated by *post-hoc* explanation methods.

2.1.2 Background

Recurrent Neural Networks (RNNs) were the pioneering attention-based structures. The work described in [48] is perhaps the most classic example of the design and implementation of computational attention mechanisms with RNNs. In the original paper, the authors focused on the task of NMT and proposed an RNN-based Encoder-Decoder architecture in which the task of the Encoder is to compress the input data (in this case, an English sentence), while the task of the decoder is to receive and decode this data to achieve a meaningful translation (in this case, a French sentence). The principal motivation for introducing an attention mechanism in this learning pipeline is related to the difficulty of achieving a perfect alignment between the input and output sequences, which is an almost impossible goal in most cases. The role of the attention mechanism is to learn weights that make the decoder focus on the relevant parts of the input sequences to produce meaningful outputs. At the time of publication, this methodology improved the state-of-the-art and paved the way for new research questions and exploration of novel attention-based frameworks.

Figure 2.1 depicts an attention-based Encoder-Decoder network [49], composed of: i) input sequences, the input vectors of the model; ii) output sequences, the output vectors of the model; iii) candidate states, the hidden states of the Encoder ("Encoder Vector"); and iv) query states, the hidden states of the decoder ("Decoder Vector"). Having defined these concepts, attention mechanisms can now be grouped into different (non-mutually exclusive) categories [49], according to the number of:

- Abstraction levels: refers to the number of representation levels where attention weights are computed. This includes single-level attention (weights are derived from the original input sequence), and multi-level attention (attention is applied across multiple abstraction levels, typically in a sequential manner).

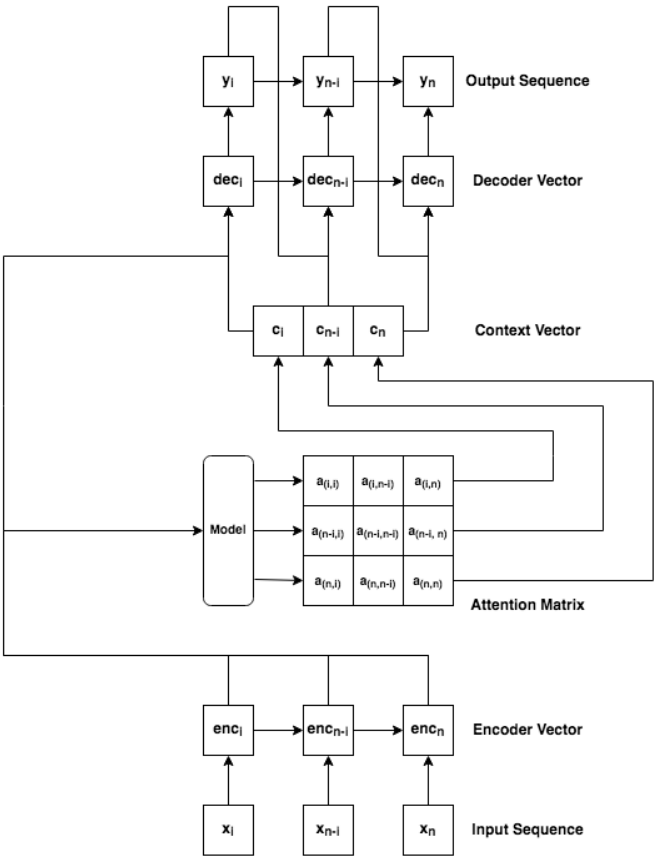


Figure 2.1: Block diagram of an attention-based Encoder-Decoder network, according to [49].

- Positions: refers to the number of input positions of the input sequence considered when computing attention. This includes soft attention (computes a weighted average of all hidden states to form the context vector), and hard attention (samples hidden states stochastically).
- Representations: refers to the number of feature representations used to compute attention. Single-representational attention learns weights from a single representation, whereas multi-representational attention combines multiple representations of the same input sequence. Multi-dimensional attention further computes relevance across different feature dimensions.
- Sequences: refers to the relationship between input and output sequences. Distinctive attention (also known as cross-attention) operates when query and candidate states belong to different sequences, co-attention jointly learns attention across multiple input sequences, and self-attention computes attention within the same sequence.

2.1.3 Attention Mechanisms for Computer Vision

Attention mechanisms for CV got their inspiration from the human visual system, which can detect whether certain features of an image are relevant or not [50]. According to Guo *et al.* [51] the different types of attention mechanisms for CV can be organized in the following categories:

- Channel attention: in deep Convolutional Neural Networks (CNNs), different channels often represent distinct semantic features [52]. Channel attention adaptively reweights channels, determining *what* information is most relevant.
- Spatial attention: this mechanism adaptively reweights spatial regions of the feature map, enabling the model to focus on *where* important information is located.
- Temporal attention: designed for data with temporal structure (e.g., video), temporal attention assigns adaptive weights to time steps, determining *when* to focus.
- Branch attention: used in multi-branch architectures, this mechanism adaptively weights different branches, determining *which* feature pathway is most relevant.
- Channel and spatial attention: combines channel and spatial mechanisms to jointly determine *what* and *where* to attend.
- Spatial and temporal attention: integrates spatial and temporal mechanisms to capture *where* and *when* relevant information occurs.

Given the scope of this thesis, we focus on channel, spatial, and self-attention, presenting only the mechanisms that were further investigated in the experimental part of this work (Section 2.1.5). For an extensive review of attention mechanisms in CV, we refer the reader to the original publication [6].

With respect to channel attention, Hu *et al.* [53] are among the first to have proposed a channel attention mechanism for CV, the Squeeze-and-Excitation (SE) block. It adaptively recalibrates channel-wise feature responses by explicitly modelling dependencies between channels (see Figure 2.2). This strengthens the representational power of CNNs by improving the quality of spatial encodings across the feature hierarchy.

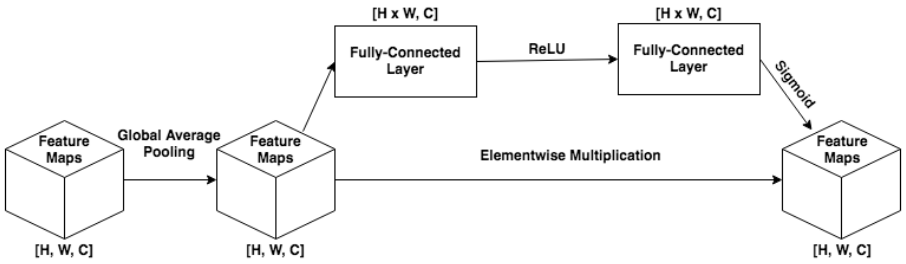


Figure 2.2: Block diagram of the SE attention block, according to [53].

With respect to channel and spatial attention, Woo *et al.* [54] have designed a Convolutional Block Attention Module (CBAM) that contains a channel attention module followed by a spatial attention module (Figure 2.3). According to the authors, this block can be easily integrated into any architecture and enables the regression of finer features. The channel attention module aims to produce a channel attention map by exploiting the inter-channel relationship of features and can be

seen as a feature detector (Figure 2.4). The spatial attention module aims to generate a spatial attention map by utilising the inter-spatial relationship of features, thus being complementary to the channel attention (Figure 2.5). Since CBAM integrates both the attention and spatial dimensions, it can be thought of as an extension of the SE block.

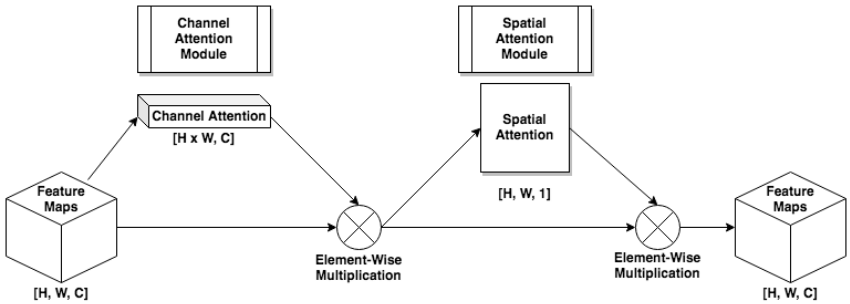


Figure 2.3: Block diagram of the CBAM, according to [54].

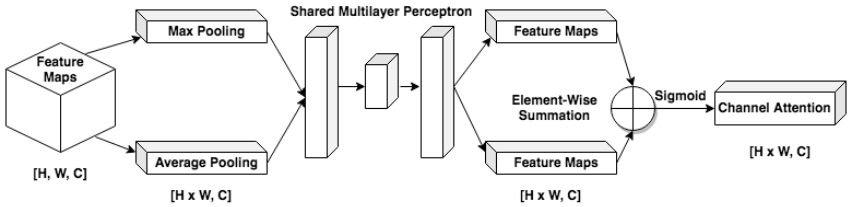


Figure 2.4: Block diagram of the Channel Attention Module of the CBAM, according to [54].

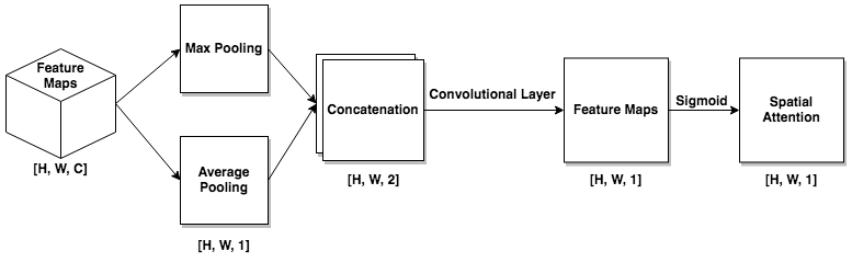


Figure 2.5: Block diagram of the Spatial Attention Module of the CBAM, according to [54].

Finally, self-attention is used in perhaps the most prominent architecture nowadays: the Transformer [46] and its respective adaptation for CV, the ViT [47]. According to Wu *et al.* [55], attention is particularly well-suited for CV tasks because: convolutions uniformly process all image patches regardless of their importance, which can lead to spatial inefficiency (e.g., image classification models should prioritize foreground over background); not all images have all concepts, therefore, applying high-level filters to all images could be computationally inefficient (e.g., features related to an image of a person may not be present in an image of a flower); and, the convolution operation is not good at establishing relationships between spatially-distant concepts, which is an essential property in CV.

2.1.4 Attention Mechanisms for Medical Image Classification

The widespread use of attention mechanisms also reached the medical domain in several tasks: classification, segmentation, report understanding, detection, reconstruction, retrieval, signal processing, and physiological and pharmaceutical research. In this thesis, we focus only on medical image classification and refer the reader to the original publication [6] for literature on the other tasks.

In breast lesion classification, Gheflati and Rivaz [56] investigated Transformer- and CNN-based architectures with data augmentation and transfer learning for ultrasound images, while Ilse *et al.* [57] introduced a gated attention mechanism in a Multiple Instance Learning (MIL) framework for histopathological image classification.

During the Coronavirus Disease (COVID)-19 pandemic, numerous studies applied attention-based models for diagnosis using Computed Tomography (CT), CXR, and ultrasound images. Several works employed Transformer-based architectures, including lightweight ViTs, hybrid Transformer–CNN models, and federated learning frameworks, to classify COVID-19 against pneumonia or healthy cases [58]–[69].

Attention mechanisms have also been applied to histopathology and whole-slide image (Whole Slide Image (WSI)) classification tasks. For example, Lu *et al.* [70] used sparse attention to aggregate patch-level features for glioma subtype classification, while other studies combined Transformer and CNN architectures to classify gastric cancer, leukemia cells, and lung cancer in WSI data [71]–[73].

In retinal disease analysis, attention-based architectures have been used for the automatic diagnosis of Diabetic Retinopathy (DR) and other ophthalmic conditions using fundus images. These approaches include gated attention mechanisms, MIL frameworks, and Transformer-based models capable of detecting lesions and performing multiclass disease classification [74]–[79].

Beyond these applications, attention mechanisms have been successfully applied to several other medical imaging tasks, including lung cancer screening, skin cancer classification, intracranial haemorrhage detection, depression recognition, multimodal MRI classification, and Alzheimer's disease diagnosis [80]–[86].

Overall, these studies showcase the widespread usage of attention mechanisms in medical imaging classification, encompassing both CNN- and Transformer-based architectures.

2.1.5 Case Studies

Most of the works on attention-based architectures for medical image classification focus solely on predictive performance. In this work we aim to go beyond this aspect by also providing insights on the impact of attention mechanisms in model complexity and interpretability. Therefore, we design an experimental set of case studies, involving three different use cases (i.e., breast lesion, chest pathology, and skin lesion) and, besides assessing the impact of attention mechanisms from the perspective of predictive performance, we also investigate their impact on interpretability, by analysing post-model attribution maps obtained with different interpretability frameworks. We

acknowledge that the visual assessment of these attribution maps is subject to a high degree of subjectivity. However, we argue that the comparative analysis of the maps obtained for models with and without attention can serve as an indirect evaluation of their degree of interpretability.

2.1.5.1 Data

To ensure that our conclusions are independent of the data, we selected medical image classification databases related to three different use cases: DR, chest pathologies and skin lesion classification. Table 2.1 presents the number of training, validation and test examples for each dataset, and the meaning of the labels we used in this classification case study.

APTOS2019 The APTOS2019 dataset contains retinography images taken under a variety of imaging conditions (mydriatic vs. non-mydriatic) and from different camera models and ethnicities. The dataset was generated by Aravind Eye Hospital in India in cooperation with Asia Pacific Tele-Ophthalmology Society (APTOS). Each retinal image has been graded independently by two ophthalmologists, with any disagreements adjudicated by a third ophthalmologist. A retinopathy severity score was assigned to each image according to the NHS diabetic eye screening guidelines[1] as R0 (no DR), R1 (background), R2 (pre-proliferative), R3 (proliferative), M0 (no visible maculopathy), M1 (maculopathy). A curated version of this dataset was used in the APTOS 2019 Blindness Detection hosted on Kaggle[2]. For the binary classification case, we grouped all the non-R0 classes into the "DR" class (see Table 2.1).

ISIC2020 The International Skin Imaging Collaboration (ISIC) 2020 [12] dataset contains dermoscopic training images of unique benign and malignant skin lesions from over 2000 patients. Each image is associated with one of these individuals using a unique patient identifier. All malignant diagnoses have been confirmed via histopathology, and benign diagnoses have been confirmed using either expert agreement, longitudinal follow-up, or histopathology. Images were collected by the following sources: Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School. The dataset was curated for the SIIM-ISIC Melanoma Classification Challenge hosted on Kaggle[3].

MIMIC-CXR The Medical Information Mart for Intensive Care (MIMIC)-CXR Database [13] is a large publicly available dataset of chest radiographs in the Digital Imaging and Communications in Medicine (DICOM) format with free-text radiology reports. The dataset contains 227835 imaging studies for 65379 patients of the Beth Israel Deaconess Medical Centre Emergency Department (Boston, MA) between 2011–2016. Each imaging study can contain one or more images,

[1] https://www.gov.uk/government/collections/diabetic-eye-screening-commission-and-provide

[2] https://www.kaggle.com/c/aptos2019-blindness-detection

[3] https://www.kaggle.com/c/siim-isic-melanoma-classification/overview

usually a frontal and a lateral view. A total of 377,110 images are available in the dataset. Studies are made available with a semi-structured free-text radiology report that describes the radiological findings of the images, written by a practicing radiologist during routine clinical care. The dataset is de-identified to satisfy the US Health Insurance Portability and Accountability Act (HIPAA) of 1996 Safe Harbor requirements. Protected health information has been removed. For the binary classification case we used a subset of images of the anterior-posterior view related to the presence or absence of pleural efusion (see Table 2.1).

Table 2.1: Data splits and label meaning for each dataset.

Dataset	Splits			Labels	
	Train	Validation	Test	0	1
APTOS2019	2334	778	550	Normal	DR
ISIC2020	19593	6568	6965	Benign	Malign
MIMIC-CXR	61203	534	1072	Normal	Pleural Efusion

2.1.5.2 Experimental Setup

We divided the experimental part of this work into several phases, described below. We conducted all the experiments using the PyTorch library [87] for Python.

Backbone Models In the first phase of this study, we trained two DL backbones (i.e., DenseNet-121 [88], ResNet-50 [89]) on each dataset. DenseNet-121's architecture [88] allows connecting all layers (with matching feature-map sizes) directly with each other, thus improving the flow of information and gradients throughout the network, and facilitating their training. ResNet50's architecture introduced the deep residual learning framework, which consists of adding skip connections that perform identity mapping, and adding their outputs to the outputs of the stacked layers. We refer the reader to the original papers for more details. We refer to these models as DenseNet-121 and ResNet-50.

Squeeze-and-Excitation Attention Block In the second phase, we adapted the SE attention block [53] to each of the backbones and trained these new architectures on the three datasets. Since one of the original implementations uses the ResNet-50, we used the author's implementation in our study, which consists of integrating the SE block between the residual layers and the activation function (in this case, the Rectified Linear Unit (ReLU)). Following this rationale, in the case of DenseNet-121, we integrated the SE block between the dense and the transition blocks. We refer to these models as SEDenseNet-121 and SEResNet-50.

Convolutional Block Attention Module In the third phase, we adapted the CBAM [54] to each of the backbones and trained these new architectures on the three datasets. Like the SE attention block, we integrated the CBAM at the same topographical location of the architectures of the backbone models. We refer to these models as CBAMDenseNet-121 and CBAMResNet-50.

Data-efficient image Transformer In the fourth phase, we tested a Transformer-based architecture, the Data-Efficient Image Transformer (DeiT) [90]. In this case, we used the DeiT-Ti version [90], which has a comparable number of parameters against the chosen CNN backbones. We refer to this model as DeiT.

Post-model Interpretability In the fifth phase, we generated visual explanation maps (i.e., saliency maps) using the Deep Learning Important Features (DeepLIFT) [91] and the Layerwise Relevance Propagation (LRP) [92] *post-hoc* methods. DeepLIFT [91] compares the activation of each neuron to its related reference activation, and assigns contribution scores according to the difference. LRP [92] is a methodology that aims to create visualisations of the contributions of single pixels to predictions. We decided to employ this strategy as a proxy for the degree of interpretability of models, thus allowing our work to be comparable to current research. We used the Captum [93] library for Python to generate these *post-hoc* saliency maps for the backbones and their attention-based versions. Regarding the DeiT, we refer the reader to [94], which proposed a framework to generate LRP attributions for Transformer-based architectures.

Data Processing Regarding the data processing pipeline, all the images were resised to the dimensions of 224×224, and a z-normalization was applied to each RGB channel. The mean and standard deviation parameters used in this operation depend on the initialization of the weights of the model. In the case of DenseNet-121, ResNet-50, SEDenseNet-121, SEResNet-50, CBAMDenseNet-121 and CBAMResNet-50, the weights were initialised from a pre-training on the ImageNet database [31] with mean $= [0.485, 0.456, 0.406]$, and std $= [0.229, 0.224, 0.225]$. In the case of DeiT, the weights were initialised from a pre-training on the ImageNet database with mean $= [0.500, 0.500, 0.500]$, and std $= [0.500, 0.500, 0.500]$.

Data Augmentation Regarding the data augmentation strategy, we decided to create a pipeline composed of several random affine transformations, i.e., random rotations, random translations, random scaling and random horizontal flip. Table 2.2 presents the parameters of data augmentation used in our experiments.

Table 2.2: Data augmentation parameters used in all experiments.

Parameter	Value
Angle of rotation in degrees	$[-10, 10]$
Horizontal translation shift	0.05
Vertical translation shift	0.1
Scaling factor	$[0.95, 1.05]$
Horizontal flip probability	0.5

Training We trained all the models in the different databases using the same optimization hyperparameters. During training, we saved the best model weights according to a decrease in the

validation loss. Table 2.3 presents the optimization hyperparameters used in our experiments. We also conducted a low data regime experiment, hence, all the models were trained using 1%, 10%, 50% and 100% of the training data. In imbalanced datasets (i.e., ISIC2020, MIMIC-CXR) we applied class-weights in the loss function.

Table 2.3: Set of hyperparameters used in all experiments. Please note that we present the batch size as a closed interval since we had to adapt its value during some of the training runs due to computational constraints.

Hyperparameter	Value
Number of epochs	300
Loss function	Cross-entropy
Optimiser	Adam
Learning rate	1×10^{-6}
Batch size	$[2, 32]$

2.1.5.3 Results and Discussion

Predictive Performance Figures 2.6, 2.7 and 2.8 respectively present the predictive performance (i.e., accuracy) results of the different models on the APTOS2019, ISIC2020 and MIMIC-CXR datasets, using different percentages of training data.

All the experiments related to the predictive performance of DL models on the different datasets suggest that it is not clear that one should expect improvements in their accuracy when using attention mechanisms. Several attention-based architectures tend to lower their predictive performances against their baseline architectures. Besides, even when directly comparing a Transformer-based architecture (i.e., DeiT), the same baseline architectures seem to achieve comparable results. Given that the intuition behind attention mechanisms is that these end up learning the most relevant features, one might expect that attention-based architectures would perform better when trained in low data regimes. However, results obtained in all datasets suggest that this might not be the case. In fact, we can observe that, apart from the Transformer-based architecture, the baseline backbone models often perform better. Hence, it is reasonable to conclude that the use of attention mechanisms will not always bring benefits to the training of DL algorithms, an argument that contrasts with a trending narrative in most recent papers. On the other hand, the results reported in the literature often relate to marginal or residual improvements in the state-of-the-art backbone networks. Given that the training of DL algorithms is generally a stochastic process, there is a need to assess these reported improvements with a more critical view and with robust statistical tests.

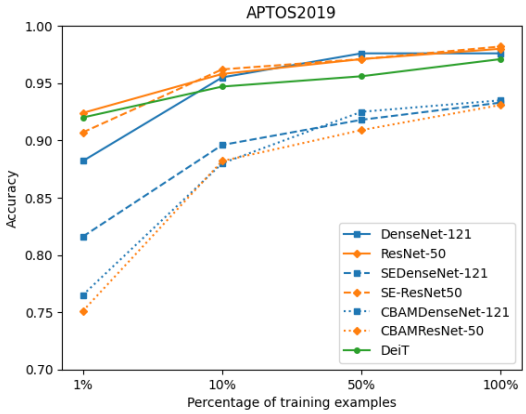


Figure 2.6: Accuracy on the APTOS2019 dataset, with different percentages of training data.

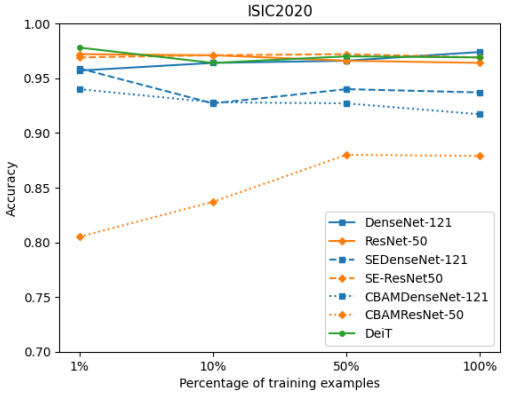


Figure 2.7: Accuracy on the ISIC2020 dataset, with different percentages of training data.

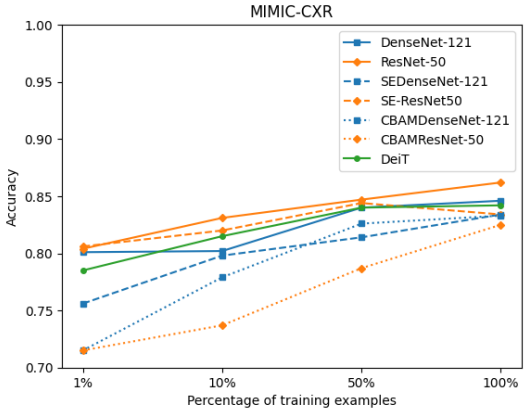


Figure 2.8: Accuracy on the MIMIC-CXR dataset, with different percentages of training data.

Model Complexity Table 2.4 presents the model complexity (i.e., the number of parameters) information for the models used in our experiments.

The integration of attention mechanisms increases the number of parameters of the DL models, thus increasing their complexity. This information allows us to conclude that, at least for CV applications, it is not necessarily true that the use of attention mechanisms contributes to the de-

crease of model complexity. While this may be true in several Natural Language Processing (NLP) applications, it is not correct to generalize such a rule to other types of applications. On the other hand, since these attention mechanisms often rely on simple operations (e.g., matrix multiplications), such as convolutions, we acknowledge that their use may reduce the training time of DL algorithms (similarly to what happened when the community started using CNNs). This aspect, though, was not objectively assessed in our experiments. However, since they are particularly efficient at modelling long-range dependencies (e.g., Transformer-based architectures), we expect these mechanisms to keep increasing their popularity and, hence, become widely used in medical image applications. Besides, Cordonnier *et al.* [95] reported that these mechanisms may operate similarly to CNNs. Another question arises from these results: are these attention-based algorithms allegedly performing well because of the inner-functioning of their attention mechanisms themselves or just because we are increasing the number of model parameters? While one may report that this issue is nonsense, we point out that some Transformer-based architectures have a considerably high number of parameters [90].

Table 2.4: Parameter count of all tested models.

Model	Number of parameters
DenseNet-121	7,054,210
SEDenseNet-121	7,357,314
CBAMDenseNet-121	7,360,706
ResNet-50	23,512,130
SEResNet-50	26,027,074
CBAMResNet-50	26,044,722
DeiT	5,486,786

Post-Model Interpretability Figure 2.9 presents examples of LRP and DeepLIFT *post-hoc* saliency maps obtained for images of the APTOS2019 dataset with labels 0 and 1 correctly classified, and with label 0 incorrectly classified by all models, respectively. The same structure is followed for examples of the ISIC2020 (Figure 2.10) and MIMIC-CXR (Figure 2.11) datasets, with the addition of examples from the label 1 incorrectly classified by all models. We used the Matplotlib [96] library for Python to generate these saliency maps visualisations, using the "bwr" colour map, which represents negative values in blue, zero values in white and positive values in red.

Although there are already several works that criticize the subjectivity of *post-hoc* saliency maps [7], [8], [97]–[100], it is still common to use these methods to justify improvements related to the degree of interpretability of models. In this sense, when analysing the results obtained with the DenseNet-121 and ResNet-50 models, we would expect that, with the increase in complexity of the attention mechanism, the distribution of the important pixels around the image would also be more focused (i.e., would have less variance). However, that does not seem to happen in our use cases. Interestingly, besides the inherent properties of the data and the task (i.e., images of the retina are different from skin images and CXR images), the results drastically change when

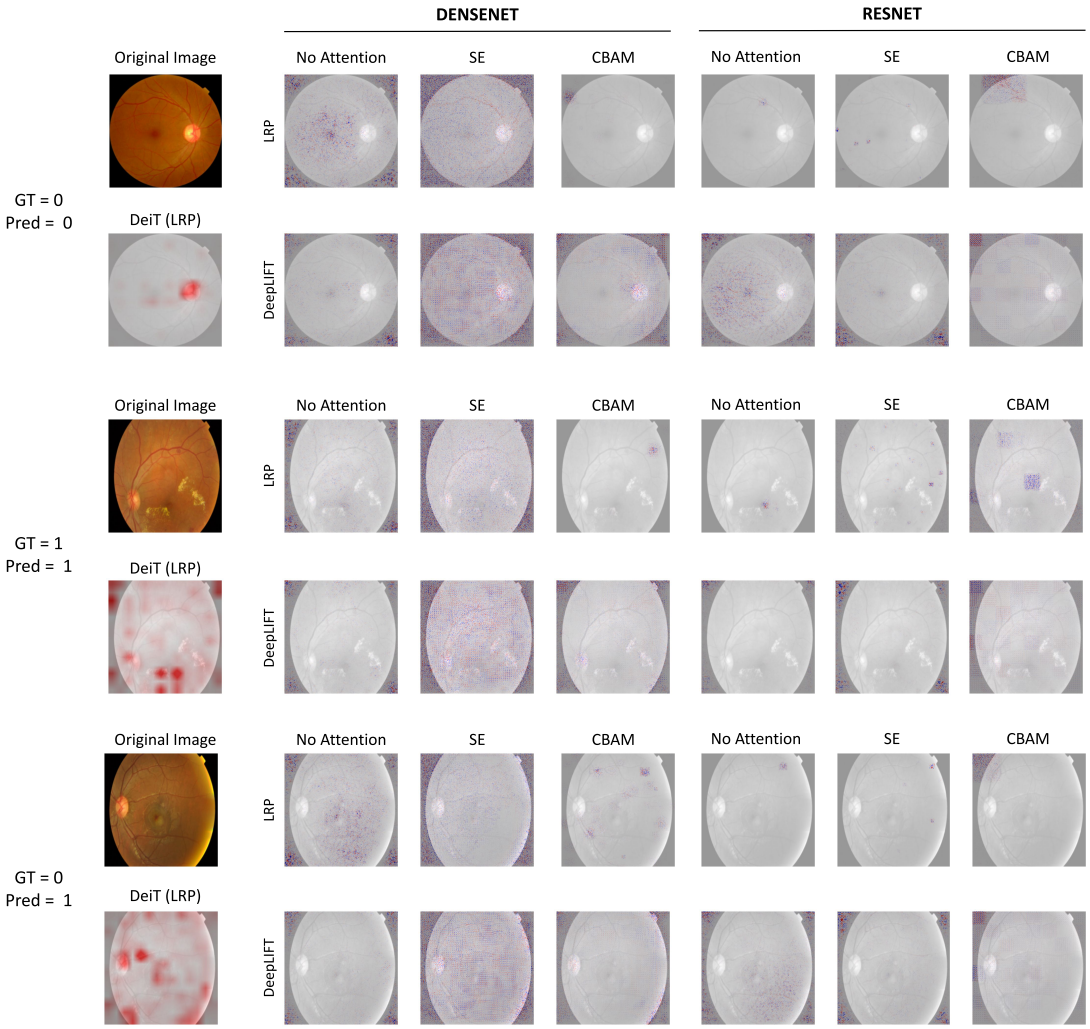


Figure 2.9: Examples of LRP and DeepLIFT *post-hoc* saliency maps for images of the APTOS2019 dataset.

we use a different backbone. In general, ResNet-based models attain less noisy DeepLIFT and LRP saliency maps, with some exceptions in the ISIC2020 dataset (e.g., Figure 2.10 when GT = 1 and Pred = 0). On the other hand, the LRP saliency maps obtained for the DeiT seem to highlight somewhat clear regions of the images. Besides, it is also important to remind the reader that these frameworks allow us to generate explanations even for the cases where the model miss-classifies. We also stress that one of the limitations of such analysis is that there does not exist an objective GT of what a high-quality visual explanation is. Therefore, it is not trivial, in CV tasks such as this, to conclude with complete confidence that, even for the cases where the model succeeds, it learned the right correlations. Hence, can we believe the narrative that attention mechanisms are learning the most relevant features of the image? Or is this type of analysis a result of luck in most cases? Truth is, our results still suggest that there is a high degree of subjectivity (i.e., the interpretation of these saliency maps deeply depends on the human that is interpreting them) and

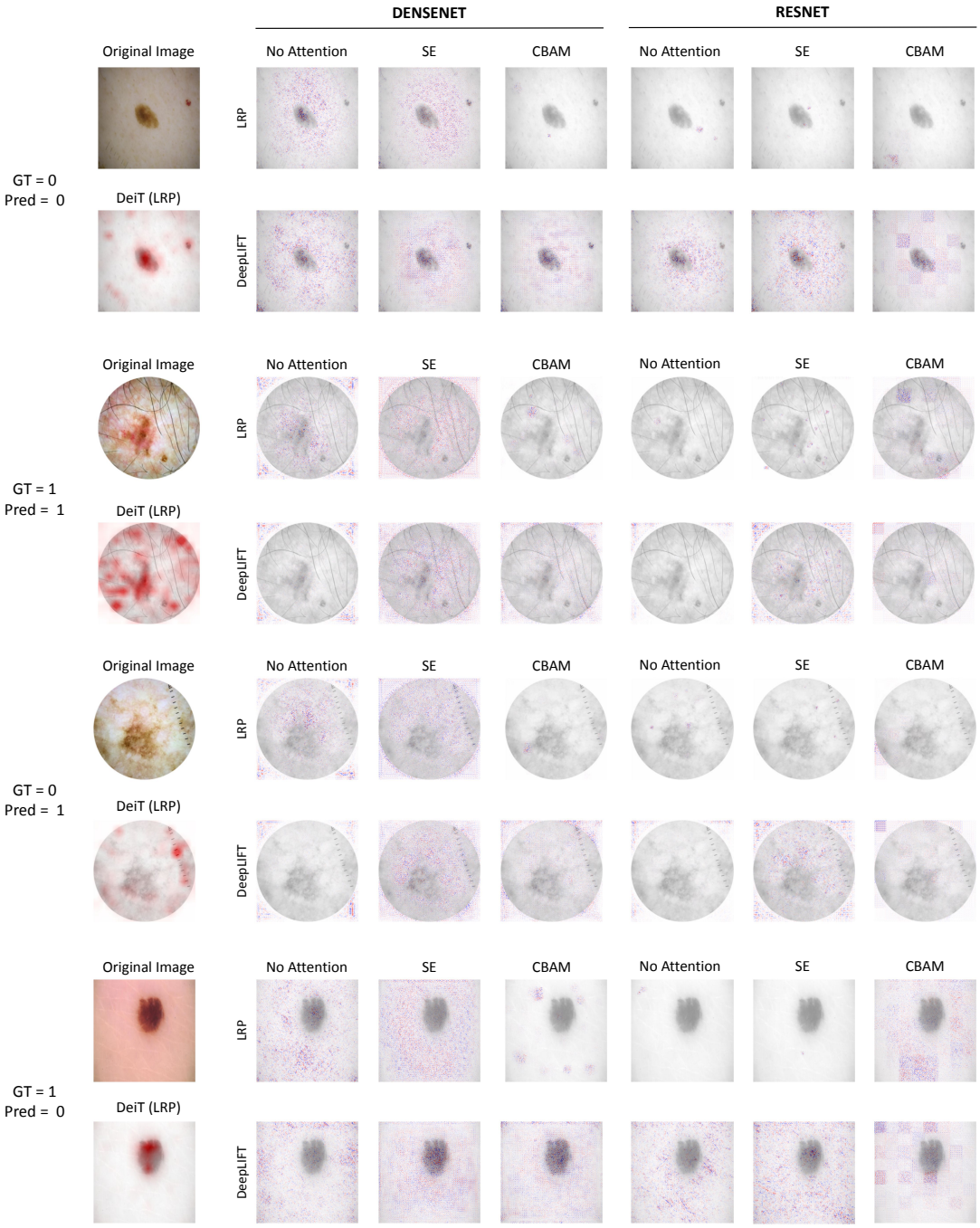


Figure 2.10: Examples of LRP and DeepLIFT *post-hoc* saliency maps for images of the ISIC2020 dataset.

that there is no apparent correlation between the use of attention mechanisms and the visual aspect of the attributions. Besides, we stress that these visualisations shown in the literature depend on several parameters (e.g., type of colour map, overlay parameter). Moreover, the method we use to generate such visualisations may often change what we expect users to observe. This is also related to another open challenge in *post-hoc* explanation methods that is the sign of the

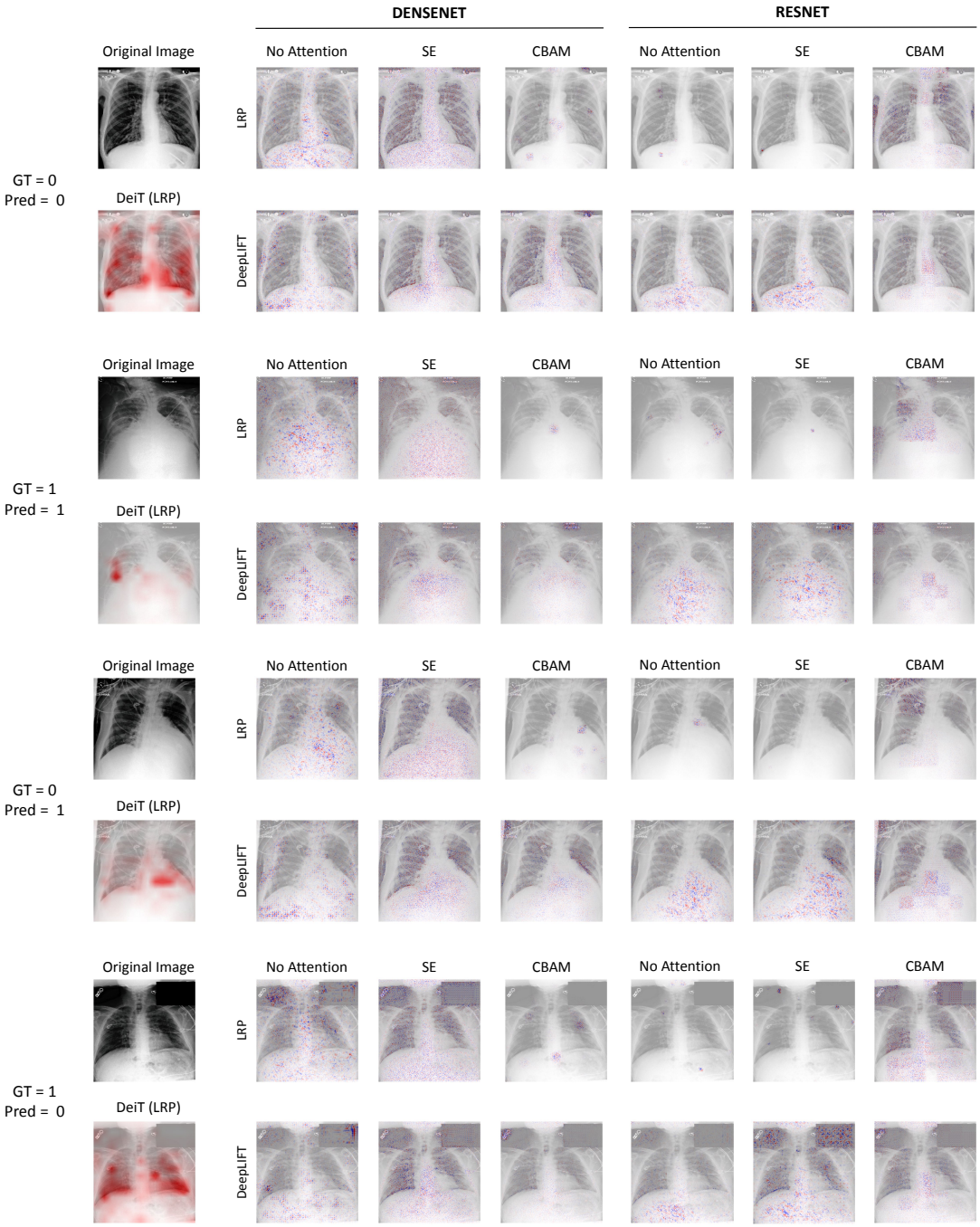


Figure 2.11: Examples of LRP and DeepLIFT *post-hoc* saliency maps for images of the MIMIC-CXR dataset.

attributions. What should we expect? Positive attributions for the positive class and negative attributions for the negative class, or the other way around? Should we always normalize the sign of the attributions so it is always positive? This motivates our statement on the need for more objective methods to assess the degree of interpretability of models.

2.1.6 Conclusions and Future Challenges

This section summarizes the main conclusions of this work and points to future directions toward the study of attention-based algorithms for medical applications.

2.1.6.1 Conclusions

This work presented an overview of the use of attention mechanisms for medical image classification and an experimental study on three different use cases. We found that backbone models can attain equivalent predictive performances to Transformer-based architectures with equivalent model complexity. Moreover, when using a *post-hoc* framework to visually assess what type of features these models can extract, we can conclude that there is still a high degree of subjectivity in such analysis. The results are very noisy, even for the cases of attention mechanisms, which is counter-intuitive (i.e., with attention mechanisms, these saliency maps should have less noise). The community is moving toward using attention mechanisms (especially Transformer-based ones) and arguing that these frameworks increase the quality, transparency, and interpretability of DL architectures. However, we state that this is not true and that there are several open challenges one needs to address to achieve this.

2.1.6.2 Future Challenges

Attention Mechanisms: Past or Future? The research on attention-based models is in constant evolution and requires a validation and maturation step. Besides, as these models keep rising in popularity and keep convincing the community of their benefits, we acknowledge that they will appear as a common block in future architectures. However, the scientific community must keep in mind that high-stake decision areas (e.g., finance, healthcare, justice) require algorithms to be fair and transparent [7]. Therefore, even if attention mechanisms are pushing DL algorithms towards the limits of their predictive power, we must start thinking about creating interpretable frameworks that allow us to audit and assess these algorithms concerning the specific conditions of their domains (i.e., in the case of this work, healthcare).

Design and Integration of Attention Mechanisms A potential disadvantage of the investment in attention mechanisms is related to the design of novel strategies of attention and their integration into well-studied backbone models. Several works on attention often rely on a specific backbone architecture for their experiments. Besides, if we look at the topographies of these DL algorithms, it is not always clear for the users where they should place these modules, and why it makes sense to put them in a specific place. Once again, another question arises: are these attention modules dependent on the backbone into which they are integrated to? Can we consider that the reported improvements related to the predictive performance of DL models are solely due to the addition of an attention module? Can we come up with an objective strategy that guides future users in building and integrating attention mechanisms into their models?

The Rise of Transformers Apart from the current deep algorithms with attention mechanisms, it is crucial to talk about Transformer-based architectures. The attentive reader might have noticed that most of the recent approaches to the use of attention-based models in medical applications rely on Transformer-based modules. While there is hype on the use of these structures, it is not clear whether they are more interpretable or not, or if their generalization power is superior to the other deep models.

Interpretability is the Path to Better Algorithms Even if we acknowledge that the Transformer-based algorithms are provoking a shift in the paradigm, it is not clear that they are improving the transparency of algorithms. For instance, in [90], the authors show that when training a DeiT using knowledge distillation techniques, the Transformer ends up learning better when the teacher model is a CNN. Hence, if the Transformer is learning with a non-interpretable model by design, how can we trust that the Transformer is inherently interpretable? On the other hand, given the capacity of these attention-based models to learn long-range dependencies, we must create objective techniques to assess the quality of the features learned by these frameworks, while moving towards the design of intrinsically interpretable attention-based architectures. Even if we intend to keep using visual saliency maps to explain our models, in a high stake decision field such as healthcare, we must achieve a clear standard, validated by the clinical community, of what these maps should look like and what is their effective meaning.

2.2 B-cos Networks

2.2.1 Context and Motivation

As shown in the previous section, post-model visual explanations can be unreliable. Furthermore, since they are computed after the MBE is trained, there are no guarantees that they are faithful to the model's decisions, with some authors arguing that in high-stakes decision fields like medicine only inherently interpretable models might guarantee trust and transparency [7]. Moreover, the XAI field lacks a gold-standard metric for evaluating the quality of explanations, so it is unclear if other methods might be better suited, especially to medical applications.

Recently, Böhle *et al.* [4] proposed a novel inherently interpretable DNN, the B-cos network, that induces an alignment between the network's weights and task-relevant input patterns through the replacement of all linear transforms by the B-cos transform. Given that the original work applied B-cos networks to natural images, achieving promising results both in terms of predictive performance and explanation quality, we investigate the application of B-cos networks to the medical domain. Thus, we test the B-cos network in several medical use cases (skin lesions, diabetic retinopathy, cervical cytology, and chest X-rays) and multiple tasks (multiclass and multilabel classification), therefore expanding the proposal of the original paper (i.e., binary classification on natural images). Given the absence of a universally accepted metric for explanation quality, we evaluate a set of properties of the generated visual explanations using an independent framework [101]. Finally, we integrate the B-cos layer into a state-of-the-art custom model for the medical domain [102], and investigate the effect of different pretraining strategies on the performance of B-cos networks.

2.2.2 Methodology

2.2.2.1 B-cos Networks

B-cos networks [4] are a type of inherently interpretable DNN, proposed by Böhle *et al.*, in which the main goal is to obtain a model where its weights are aligned with task-relevant input patterns. This is done by replacing linear layers that compute the dot product between the input vector $\mathbf{x}$ and the weight vector $\mathbf{w}$ (see Equation 2.1) with B-cos layers that scale $cos(\angle(\mathbf{x},\mathbf{w}))$ by raising it to a power of $B-1$ (see Equation 2.2).

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\|c(\mathbf{x},\mathbf{w}) \tag{2.1}$$

with $c(\mathbf{x},\mathbf{w}) = cos(\angle(\mathbf{x},\mathbf{w}))$.

$$\text{B-cos}(\mathbf{x};\mathbf{w}) = \widehat{\mathbf{w}}^T\mathbf{x} \times |cos(\mathbf{x},\widehat{\mathbf{w}})|^{B-1} \tag{2.2}$$

where the hat-operator scales $\widehat{\mathbf{w}}$ to have unit norm, and ($B \geq 1$) is a hyperparameter controlling how strongly the model favours aligned features.

When $B = 1$, the transformation reduces to the standard linear operation, but for larger B the output increasingly emphasizes inputs whose direction closely matches the learned weights. This alignment property ensures that predictions depend primarily on features that are directionally consistent with the learned representation, making it possible to express the network's decision as an input-dependent linear combination of features. Consequently, the contributions of individual input components can be directly interpreted, enabling faithful and intrinsic explanations of model predictions without relying on external attribution methods.

Consider a B-cos network f^* with layers $\mathbf{l}_j^*$ and parameters θ. The network can be written as:

$$f^*(\mathbf{x};\theta) = \mathbf{l}_L^* \circ \mathbf{l}_{L-1}^* \circ \mathbf{l}_2^* \circ \mathbf{l}_1^*(\mathbf{x}) \tag{2.3}$$

where each B-cos layer j with input $\mathbf{a}_j$ and weight $\mathbf{W}_j$ is given by:

$$\mathbf{l}_j^*(\mathbf{a}_j;\mathbf{W}_j) = \widetilde{\mathbf{W}}_j(\mathbf{a}_j)\mathbf{a}_j \tag{2.4}$$

with

$$\widetilde{\mathbf{W}}_j(\mathbf{a}_j) = |c(\mathbf{a}_j;\widehat{\mathbf{W}}_j)|^{B-1} \odot \widehat{\mathbf{W}}_j \tag{2.5}$$

Thus, according to Equation 2.3, the network can be represented as:

$$f^*(\mathbf{x};\theta) = \widetilde{\mathbf{W}}_L(\mathbf{a}_L)\widetilde{\mathbf{W}}_{L-1}(\mathbf{a}_{L-1})...\widetilde{\mathbf{W}}_1(\mathbf{a}_1 = \mathbf{x})\mathbf{x} \tag{2.6}$$

$$= \mathbf{W}_{1\to L}(\mathbf{x})\mathbf{x} \tag{2.7}$$

with

$$\mathbf{W}_{1\to L}(\mathbf{x}) = \prod_{j=1}^{L} \widetilde{\mathbf{W}}_j(\mathbf{a}_j) \tag{2.8}$$

In other words, since B only rescales the dot product based on the alignment (i.e., the cosine) between the input and the weights (see Equation 2.2), the entire network can be mathematically rewritten for a specific input as a single linear model (see Equation 2.7). This means the final class score can be expressed as a weighted sum of the input features, where the weights depend on how the intermediate representations align with the learned filters. To compute an explanation, the model propagates these alignment-based scaling factors through the network and multiplies them with the original weights, producing an effective weight vector for the input (see Equation 2.8). The contribution of each input feature (for example, each pixel in an image) is then obtained by multiplying that feature with its corresponding effective weight, producing a contribution map that highlights which regions positively support the predicted class. Because these contributions come from the model's exact computation rather than an external approximation, the resulting explanations are, by design, faithful to the model's decision process.

2.2.2.2 Explanation Quality Assessment Metrics

Recognising the importance of a proper methodology to assess the quality of XAI methods, we decided to evaluate the approach proposed by Böhle *et al.* [4] with the framework developed by Hedström *et al.* [101]. It consists of a battery of quantitative measurements of different properties that a good explanation should have. We consider the following properties and metrics [101]:

- Faithfulness (do features deemed relevant affect model decisions more strongly?): Selectivity [103], [104] (drop in performance when features deemed relevant by a XAI method are iteratively removed) and Faithfulness Correlation [105] (correlation between the sum of the subset of randomly replaced attributions and the difference in function output)
- Robustness (are explanations stable when subject to slight perturbations in the input?): Maximum Sensitivity [106] (maximum sensitivity of an explanation using Monte Carlo sampling) and Relative Representation Stability [107] (distance between the internal model representation of an explanation and its perturbed version)
- Complexity (are explanations concise?): Sparsity [108] (Gini Index to measure if only highly attributed features are predictive of the model output) and Complexity [105] (entropy of the contribution of all features to the total magnitude of the attribution)
- Localisation (is the explanatory evidence centered around a given region of interest?): Focus [109] (quantifies the precision of the explanation by creating mosaics of images from different classes) and Attribution Localisation [110] (ratio of positive attributions within a targeted object and the total positive attributions)

2.2.2.3 Pretraining Strategies

In the work by Böhle *et al.* [4], the authors compared B-cos architectures with their original counterparts, but only on ImageNet [31]. Thus, their experiments involved training both types of networks from scratch. In the medical imaging domain, it is common practice to employ transfer learning [111]. Therefore, we test different pretraining strategies for both B-cos and original networks: no pretraining, pretraining on ImageNet, and initialising B-cos networks from differently trained baselines.

2.2.2.4 B-cos Version of a Custom Medical Imaging Model

To understand the feasibility of integrating the B-cos transform in medical DL models, we adopted the work of Liu *et al.* [102] as a use case. Their Clinical-Inspired Network (CI-Net) is an attention-based model that aims to mimic the diagnostic process of clinicians, having achieved state-of-the-art performance on 6 benchmark databases for skin lesion analysis. Contrary to commonly used networks like the DenseNet [88], the CI-Net contains two branches and a distinguish module, which learns to discriminate images from different classes, making it more difficult to integrate the B-cos transform into this architecture.

2.2.2.5 Datasets

Given the diversity of use cases and tasks in the medical domain, we test B-cos networks on different tasks (multiclass and multilabel classification), and applications (skin lesions, DR, cervical cytology, and CXRs). For skin lesion analysis, we used the ISIC2018 dataset [16], [17], containing 7 classes from melanoma to vascular lesions. For DR classification, we used the APTOS2019 dataset [11], which includes images with 5 different severity scores, ranging from no DR to proliferative DR. For cervical cytology we used the DTU/Herlev Pap Smear Database [15], consisting of images of individual cervical cells in different stages of cervical cancer. For CXR diagnosis (and multilabel classification), we resorted to the VinDr-CXR [18], which contains 22 labels and corresponding bounding boxes for pathologies like pneumonia or cardiomegaly.

2.2.2.6 Implementation Details

All the models followed the training pipelines described by Böhle *et al.* [4], except for CI-Net-based models, which followed the training pipeline described by its authors [102]. Regarding data processing and augmentation, we resised all the images to 448×448, applied random affine transforms (i.e., rotation, translation, scale, and shear) according to specific literature on each use case (more details in the code), and normalised the inputs accordingly. All datasets were divided into stratified train-val-test splits with 20% for test for datasets without test set, and 10% of the remaining data for validation. For ISIC2018, we used the official split. We saved the best models according to the best F1-score on the validation set.

For the explanation quality metrics, we ensured that all models were being evaluated under the same conditions: models were evaluated only on the set of correctly predicted images by all models under test. For the multilabel classification scenario (VinDr-CXR), an image is considered correctly classified only when all labels are correctly predicted, and we explain only the label with the lowest index. For datasets where more than 100 images are correctly predicted by all models, we evaluate only 100 (for Herlev we evaluated 40 images). More details can be found in the available code.

2.2.3 Results and Discussion

We present our results on the 4 datasets considered in Table 2.5.

2.2.3.1 Training

Except for the CI-Net, we focus on DenseNet-121 [88], since it is a popular architecture among the medical imaging community and Böhle *et al.* [4] already showed that their conclusions hold for different architectures. One important aspect to highlight is that B-cos networks take twice as long to train and require about twice as much memory.

2.2.3.2 Pretraining Strategies

For ISIC and APTOS, we study the effect of different pretraining strategies. In both cases, the performance improves on baseline and B-cos models when ImageNet pretraining is used (e.g., model 3 vs 6 or 4 vs 7). Interestingly, on APTOS2019, transferring knowledge from the baseline model already adapted to the final domain worsens the results compared to the B-cos network trained from scratch (models 13 and 16 vs 12), and, on both datasets, it falls short of the performance obtained when transferring knowledge from a B-cos network trained on another domain (models 5 and 8 vs 7, 13 and 16 vs 15). This shows that it is preferable to leverage knowledge from another B-cos network that already learned the weight-input alignment pressure, even though on another domain, than to learn from a non B-cos network (i.e., non-aligned) on the same domain.

2.2.3.3 Adapting a Custom Model

Transforming a custom model into a B-cos network is a straightforward process given the code made available by Böhle *et al.* [4]. As expected, this custom architecture outperforms DenseNet-121 on both baseline and B-cos versions in terms of Localisation. Interestingly, the baseline CI-Net is not better in terms of classification than the DenseNet-121, but the B-cos version achieves the best results (see models 9 and 10).

2.2.3.4 Explanation Quality

We started by applying our evaluation protocol to ImageNet-trained networks made available by the original authors to verify if the results they reported held for different properties of explanations (in their work [4], only the localisation property is evaluated with a metric similar to Focus). In 4 out of 7 metrics, their conclusions hold, i.e., *post-hoc* methods applied on B-cos networks yield better explanations and, in turn, the inherent B-cos explanations surpass other methods both on B-cos and non B-cos networks. On the medical use cases tested, the results vary. For ISIC and APTOS, the trend in terms of Selectivity is reversed compared to the trend on ImageNet. For Herlev and VinDr-CXR, explanations obtained from B-cos networks are better than on the original networks. Still, the inherent B-cos explanations do not surpass other XAI methods (e.g., 4.56 for Grad-Class Activation Map (CAM) vs 5.04 on VinDr-CXR). In terms of Faithfulness Correlation, B-cos explanations surpass other methods except on VinDr-CXR. Regarding robustness, B-cos networks and explanations are, in general, better for RRS but worse for Maximum Sensitivity. The same applies to Sparseness and Complexity. Finally, on Localisation metrics, B-cos explanations surpass the best explanation methods on B-cos and non B-cos models for 3 out of 4 datasets, but only for ImageNet pretrained models (c.f. models 6/7, 9/10, 14/15, 19/20). Further, on VinDr-CXR, where local bounding box annotations are available, the Attribution Localisation metric shows the trend the original authors found. These results indicate that B-cos explanations are, indeed, generally more focused than other explanations, which is also corroborated by the qualitative results of Figure 2.12. Visually, Input$\times$Gradient produces more sparse maps, while

Grad-CAM and B-cos explanations are more concentrated around a single area. Furthermore, B-cos maps tend to have less highly activated pixels than Grad-CAM's.

Table 2.5: Results on model performance and XAI quality assessment per model across several datasets. For all models we show the results for the best XAI method. For B-cos networks we also show results for the intrinsic B-cos explanations. We compare against the following XAI methods: G - Grad, GC - Grad-CAM [5], I×G - Input×Gradient [91], IG - Integrated Gradients [112]. For each metric ↑ (↓) means higher (lower) is better. Underlined numbers represent the best result for non B-cos models on a given dataset, while bold numbers highlight the best B-cos results. DN121 stands for DenseNet-121 and RN18 IN corresponds to an ImageNet trained ResNet-18.

Dataset	ID	Model	Pretraining	Task Performance			Faithfulness		Robustness		Complexity		Localisation	
				(B)ACC ↑	F1 ↑	AUC ↑	Select. ↓	Correl. ↑	Sens. ↓	RRS ↓	Sparse. ↑	Complex. ↑	Focus ↑	Att. Loc. ↑
ImageNet [31]	1	DN121	-	72.6	72.3	99.8	7.63 (IG)	0.073 (I×G)	0.545 (GC)	1.62 (G)	0.610 (I×G)	10.5 (GC)	0.722 (GC)	-
	2	B-cos DN121	-	**73.6**	**73.2**	99.2	5.92 (G) **4.85**	**0.138** (IG) 0.043	0.994 (GC) 1.54	1.22e-4 (G) **6.92e-5**	0.604 (IG) **0.624**	10.3 (GC) 10.1	0.748 (GC) **0.851**	- -
ISIC2018 [16]	3	DN121	-	76.7	62.5	92.4	25.9 (IG)	0.063 (GC)	1.80 (IG)	0.012 (IG)	0.583 (G)	11.6 (IG)	0.707 (GC)	-
	4	B-cos DN121	-	74.9	57.8	89.1	48.1 (GC) 91.4	**0.487** (I×G) -0.020	**0.964** (GC) 1.13	4.33e-6 (I×G) 8.87e-8	0.612 (I×G) 0.443	**11.9** (GC) 11.8	0.434 (GC) 0.585	- -
	5	B-cos DN121	3	75.5	59.6	90.4	35.7 (IG) 70.7	0.297 (I×G) 0.057	2.19 (GC) 1.94	2.84e-6 (IG) **4.81e-8**	0.657 (IG) 0.550	11.4 (G) 11.6	0.539 (GC) 0.614	- -
	6	DN121	1	85.6	78.5	97.6	100.3 (IG)	0.288 (GC)	1.32 (G)	0.507 (IG)	0.606 (GC)	11.7 (IG)	0.772 (GC)	-
	7	B-cos DN121	2	82.3	71.2	93.7	65.6 (IG) 46.6	0.114 (I×G) 0.024	2.44 (G) 6.62	9.31e-7 1.70e-5	0.707(GC) **0.767**	11.5 (I×G) 10.9	0.619 (I×G) 0.796	- -
	8	B-cos DN121	6	77.1	62.0	90.9	**26.8** (GC) 48.4	0.170 (I×G) 0.027	2.68 (GC) 1.62	2.58e-7 (G) 1.33e-7	0.631 (I×G) 0.616	11.5 (G) 11.5	0.668 (GC) 0.656	- -
	9	Baseline CI-Net [102]	RN18 IN	84.2	74.2	96.3	30.3 (I×G)	0.173 (GC)	2.00 (IG)	0.250 (G)	0.716 (GC)	11.4 (IG)	0.887 (GC)	-
	10	B-cos CI-Net [102]	RN18 IN	**82.9**	**71.4**	**94.4**	52.8 (I×G) 77.3	0.144 (IG) 0.061	1.40 (G) 3.51	1.55e-5 (G) 1.10e-5	0.744 (GC) 0.709	11.5 (I×G) 11.2	0.591 (IG) **0.889**	- -
APTOS2019 [11]	11	DN121	-	79.3	57.6	93.0	32.3 (I×G)	0.049 (GC)	1.49 (G)	0.038 (IG)	0.791 (GC)	11.4 (IG)	0.695 (GC)	-
	12	B-cos DN121	-	79.3	58.6	90.9	88.1 (GC) 82.3	**0.365** (IG) 0.019	29.1 (G) 18.4	1.02e-5 (I×G) 1.91e-5	0.723 (IG) 0.735	**11.4** (GC) 11.0	0.413 (GC) 0.585	- -
	13	B-cos DN121	11	75.7	46.2	88.1	**41.5** (I×G) 51.7	0.104 (I×G) 0.027	1.66 (GC) **1.46**	1.08e-6 (IG) **7.79e-7**	0.721 (GC) 0.723	11.1 (G) 11.0	0.549 (GC) 0.642	- -
	14	DN121	1	82.9	66.2	94.7	37.1 (IG)	0.180 (GC)	1.41 (G)	0.156 (G)	0.802 (GC)	11.2 (IG)	0.845 (GC)	-
	15	B-cos DN121	2	**81.3**	**61.9**	**92.6**	56.1 (I×G) 69.3	0.061 (GC) 0.034	2.61 (IG) 5.10	9.64e-7 (G) 1.35e-6	0.784 (GC) **0.826**	11.2 (I×G) 10.5	0.729 (GC) **0.870**	- -
	16	B-cos DN121	14	78.9	54.8	91.4	51.0 (IG) 53.6	0.0578 (GC) 0.046	1.61 (I×G) 4.99	1.05e-6 (G) 1.67e-5	0.792 (GC) 0.808	10.8 (I×G) 10.6	0.791 (GC) 0.818	- -
Herlev [15]	17	DN121	1	66.2	71.0	94.3	10.1 (I×G)	0.143 (GC)	0.912 (GC)	0.0530 (Grad)	0.553 (I×G)	10.6 (GC)	0.732 (GC)	-
	18	B-cos DN121	2	63.5	67.1	91.4	**8.67** (I×G) 9.65	0.124 (GC) **0.144**	2.03 (IG) 2.68	**1.29e-6** 2.24e-6	0.557 (IG) **0.572**	10.3 (GC) 10.2	0.606 (GC) 0.666	- -
VinDr-CXR [18]	19	DN121	1	96.2	28.4	74.8	9.13 (I×G)	0.119 (GC)	1.00 (I×G)	2.97e-4 (I×G)	0.740 (GC)	11.6 (IG)	0.277 (I×G)	0.137 (GC)
	20	B-cos DN121	2	95.6	22.7	**77.8**	**4.56** (GC) 5.04	0.093 (GC) 0.066	2.58 (I×G) 13.2	**1.18e-6** (IG) 1.60e-5	**0.763** (GC) 0.726	11.5 (G) 11.0	0.297 (I×G) **0.301**	0.237 (GC) **0.287**

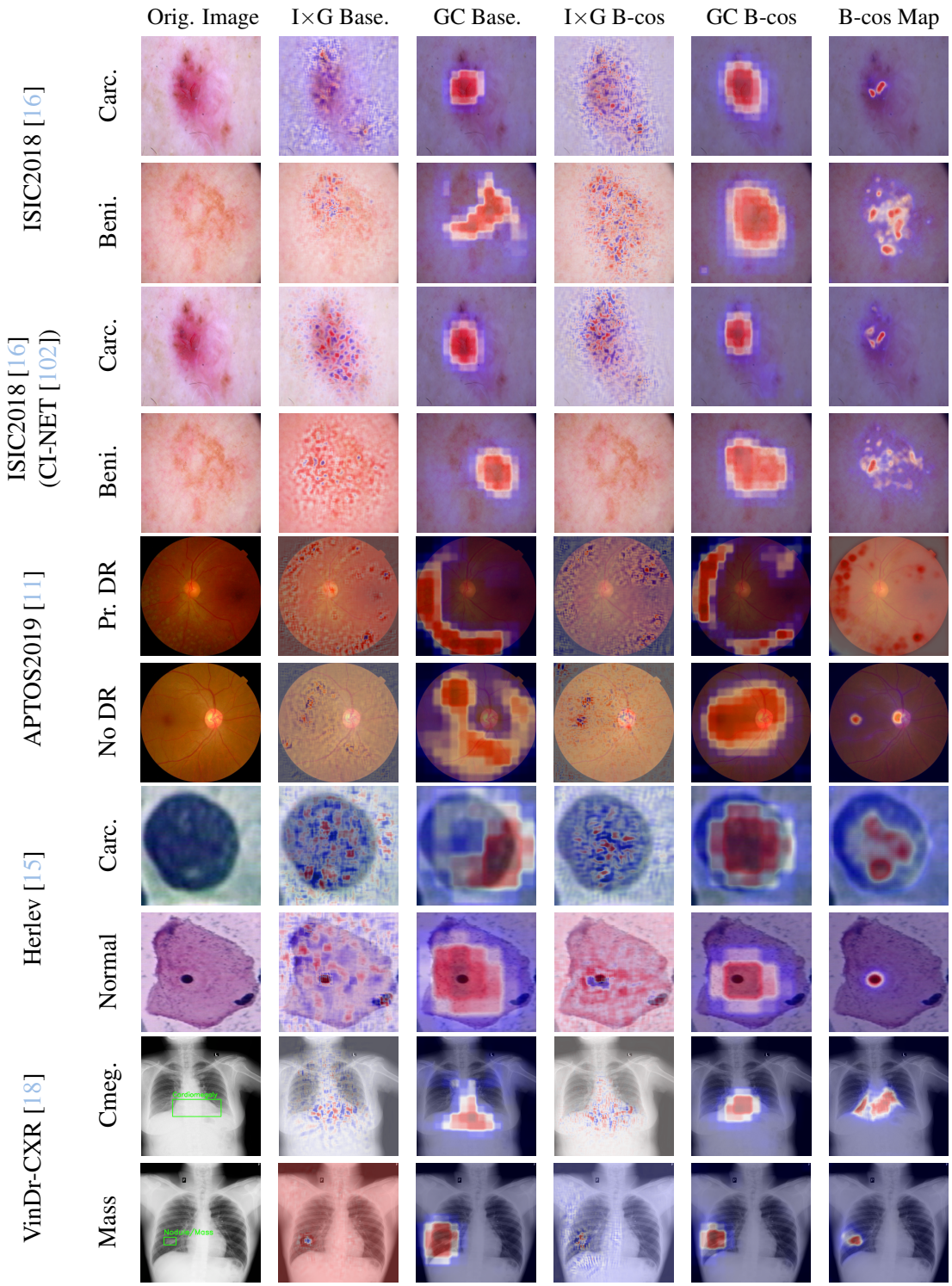


Figure 2.12: Comparison of explanations given by Input×Gradient [91] (I×G), Grad-CAM [5] (GC) and B-cos on original (Base.) and B-cos networks. The colourmap ranges from blue to red, blue(red) meaning less(more) relevant pixels.

2.2.4 Summary and Conclusions

We investigated the applicability and suitability of B-cos networks for the medical domain. Although achieving better explanation quality than other XAI methods on ImageNet trained models, it is not clear that the same happens for the medical use cases considered, especially when taking into account all explanation quality assessment metrics. However, B-cos explanations yield more localised saliency maps, a finding that corroborates the original authors' conclusions. Given these results and the slight drop in classification performance, it is not clear that B-cos networks produce better explanations in the medical domain. Further work should be devoted to integrating B-cos networks in different tasks (e.g., image segmentation), studying of different XAI quality assessment metrics with more *post-hoc* methods, and a deep focus on model optimization or regularization to promote the integration of clinical-domain knowledge into the models (e.g., integrating clinical metadata through data fusion techniques, using class weights in the loss function).

Chapter 3

Concept-based Explanations

Foreword on Author Contributions

The research work described in this chapter was conducted in collaboration with Cristiano Patrício, under the supervision of Professors Jaime S. Cardoso, Luís F. Teixeira, and João C. Neves. The author of this thesis contributed to this work equally with Cristiano Patrício, as both authors worked on every step of the paper: conceptualization, implementation, results analysis and discussion, and preparation of the scientific publication. The results of this work have been disseminated in one article in an international journal:

[20] Isabel Rio-Torto*, Cristiano Patrício*, Jaime S. Cardoso, Luís F. Teixeira, and João C. Neves, "CBVLM: Training-free Explainable Concept-based Large Vision Language Models for Medical Image Classification" Computers in Biology and Medicine, vol. 198, p. 111 145, 2025

* Equal contribution

Code: https://cristianopatricio.github.io/CBVLM/

In this chapter we move on to another explanation modality: concept-based explanations. While visual explanations can be hard to interpret, especially in contexts requiring specialised knowledge (e.g., CXR analysis), concept-based explanations are inherently interpretable, since they are composed of (predefined) sets of human-understandable concepts. Once more in accordance with one of the goals of this thesis, the method introduced in this chapter aims to increase interpretability while keeping the additional annotation burden as low as possible.

3.1 Context and Motivation

As mentioned in Chapter 1, the remarkable performance of supervised DL models greatly depends on the availability and quality of training samples, which are particularly costly to obtain, especially in the medical domain, as clinician expertise is required. LVLMs, which have recently made their way into the medical field with Med-LVLMs such as [113]–[116], are no exception. In

fact, they usually require even more training data. Consequently, a different research direction has emerged, exploring the potential of applying LVLMs to specific downstream tasks with minimal or no supervision, by leveraging their few-shot capabilities [117]–[119], particularly the use of ICL for improving model performance through a short number of examples in the prompt. In addition to the need for annotated data, another major challenge with DL-based medical imaging, and the main focus of this thesis, is the lack of interpretability. One way of introducing interpretability into such frameworks is via concept-based explanations. CBMs provide this type of explanations by introducing a layer, the "bottleneck", that predicts the presence or absence of a predefined set of human-interpretable concepts. The final classification prediction is then based on these concepts. Nevertheless, CBMs still require training and annotated data; in fact, the amount of annotated data actually increases since concept annotations are needed on top of the classification labels. We propose combining the best of both worlds: the interpretability of CBMs and the few-shot abilities of LVLMs to decrease the annotation cost. Our methodology, *CBVLM*, encompasses a two-stage approach, where, first, the LVLM is prompted to predict the existence/absence of a set of concepts in an image, and, in a second stage, the predicted concepts are incorporated into the prompt such that the LVLM provides a final diagnosis grounded on these predicted concepts. Figure 3.1 represents this methodology compared to the common approach where an LVLM is simply prompted to perform classification. Moreover, contrary to CBMs, which require retraining the entire model to incorporate a new concept, *CBVLM* can easily incorporate additional concepts, requiring only two inference steps. We validate *CBVLM* on four medical datasets encompassing different imaging modalities and using twelve LVLMs, and show that *CBVLM* achieves comparable performance or even surpasses task-specific supervised methods, doing so without training, with few annotated examples, and without sacrificing interpretability. Although, individually, the techniques used in *CBVLM* (e.g., ICL) are not novel, this is, to the best of our knowledge, the first work combining them in a novel methodology exploring the capability of LVLMs to provide concept-based explanations for medical image classification.

3.2 Related Work

3.2.1 Large Vision-Language Models in the Medical Domain

The emergent capabilities of Large Language Models (LLMs) have led to the development of LLMs able to process visual inputs, i.e., LVLMs. These models can reason about images in a zero- or few-shot manner (e.g., LLaVA [120], Generative Pre-trained Transformer (GPT)-4V [121]), allowing them to tackle downstream tasks without fine-tuning. General-purpose LVLMs have recently been adapted for the medical domain through training on extensive medical datasets, coining the term Med-LVLMs. The generalization capability of these models has prompted researchers to assess their performance in multiple medical data. Royer *et al.* [122] recently released an open-source toolkit for fair and reproducible evaluation of Med-LVLMs in tasks like classification and report generation. Han *et al.* [117] evaluated the performance of a general LVLM on the NEJM

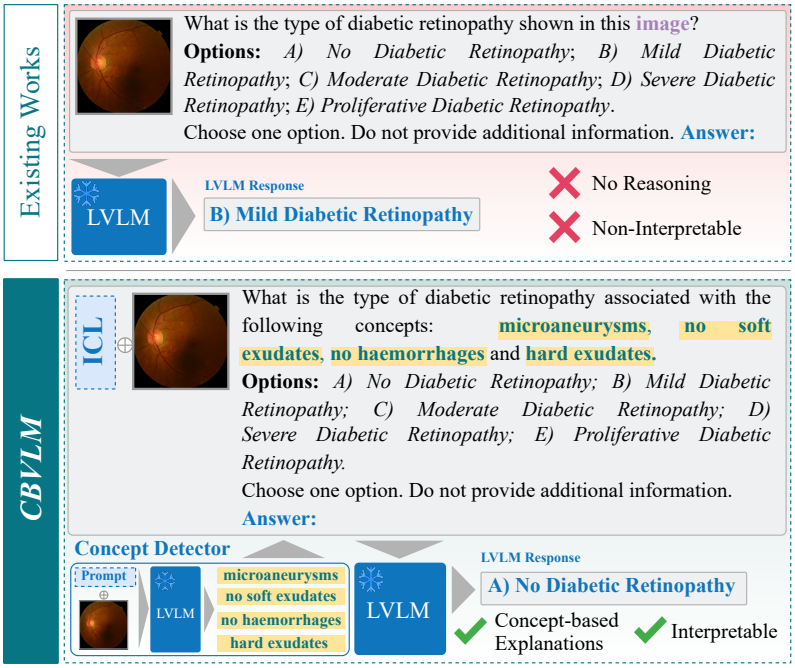


Figure 3.1: Comparison between the proposed methodology (*CBVLM*) and existing methods. Unlike traditional approaches, our methodology grounds the final diagnosis on a set of clinical concepts predicted by the LVLM itself (*cf. Concept Detector*), thus increasing classification performance, while being interpretable.

Image Challenge, as well as the impact of fine-tuning the LVLM. In [123], Med-LVLMs were evaluated in different specialities and compared with medical experts, with results comparable to the performance of junior doctors. In [119], the authors benchmarked the performance of GPT-4V against fully supervised image classifiers and found that their performance is comparable. In a different line of research, Xia *et al.* [124] evaluated the trustworthiness and fairness of Med-LVLMs and found that models often display factual inaccuracies and can be unfair across different demographic groups. *CBVLM* distinguishes itself from previous studies as it is not limited to prompting the LVLM to provide a diagnosis, but it is, to the best of our knowledge, the first to generate concept-based explanations on which the LVLM response is grounded prior to diagnosis prediction, thus enhancing the performance and transparency of the decision-making process.

3.2.2 Concept Bottleneck Models

CBMs [125] are based on an Encoder-Decoder paradigm, where the Encoder predicts a set of human-specified concepts from the input image, while the decoder leverages these predicted concepts to generate the final predictions. Due to their inherent interpretability, these models have attracted particular interest in medical image classification. Several CBMs have been proposed for a wide variety of medical tasks (e.g., melanoma diagnosis [126]–[128], knee osteoarthritis [125] and CXR classification [129]). Nevertheless, CBMs have well-known limitations: (i) they require concept annotations (image-level or class-level), (ii) they tend to have lower performance

compared to traditional (black-box) supervised methods, (iii) they struggle to predict concepts accurately, and (iv) incorporating new concepts implies retraining the whole CBMs . These challenges are particularly significant in medical imaging, where data availability is typically limited due to the expertise required for annotations. Consequently, LVLMs represent a promising alternative to traditional CBMs, as they are able to generalize to unseen domains with minimal or no supervision. Moreover, incorporating additional concepts becomes trivial and does not require fine-tuning, as happened with traditional CBMs.

3.3 Methodology

Given the task of predicting a target $y \in \mathbb{R}$ from input $\mathbf{x} \in \mathbb{R}^d$, let $\mathscr{D} = \{(\mathbf{x}^{(i)}, y^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^n$ be a batch of training samples where $\mathbf{c} \in \mathbb{R}^{\{0,1\}^l}$ is a vector of l concepts. Traditional (black-box) models learn a function $h : \mathbb{R}^d \to \mathbb{R}$ that directly predicts the target $y \in \mathbb{R}$ from the input $\mathbf{x} \in \mathbb{R}^d$. However, CBMs first map the input $\mathbf{x}$ into a set of interpretable concepts $\mathbf{c}$ ("bottleneck") by learning $g : \mathbb{R}^d \to \mathbb{R}^l$, and use these concepts to predict the target y using $f : \mathbb{R}^l \to \mathbb{R}$. Therefore, the prediction $\hat{y} = f(g(\mathbf{x}))$ is entirely based on the predicted concepts $\hat{\mathbf{c}} = g(\mathbf{x})$.

In order to generate concept-based explanations with LVLMs, we follow the CBMs paradigm and introduce a two-stage methodology, as outlined in Figure 3.2 and detailed in the following subsections.

3.3.1 Concept Detection

The first stage involves predicting the l concepts present in a given input image. For each predefined clinical concept, we start by providing the LVLM with a short description of that concept and then prompt it to answer if that concept is shown in the image (see Figure 3.2 on the left). Thus, concept prediction is given by:

$$p_{LVLM}(t_{ans} | t_I^1, ..., t_I^m, t_P^1, ..., t_P^k) \tag{3.1}$$

where t_{ans} corresponds to the first generated token, $t_I^1, ..., t_I^m$ correspond to the input image tokens, and $t_P^1, ..., t_P^k$ correspond to the tokens of the prompt (description of the concept and the question). The structure of the (zero-shot) prompt for the prediction of a single concept is as follows:

> {A brief description of the {concept}}.
>
> In the <image>, the {concept} is:
>
> A) present
>
> B) absent
>
> Choose one option. Do not provide additional information. Answer:

The descriptions for each clinical concept, as presented in the *Instruction* section shown in Figure 3.2, were generated by ChatGPT [130] with the following prompt: "*According to published*

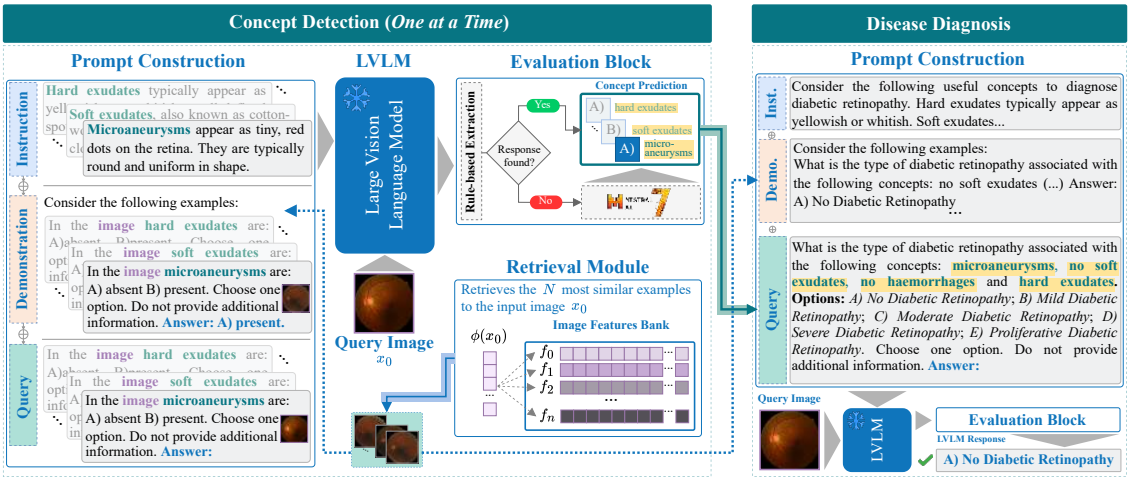


Figure 3.2: Overview of *CBVLM*. Our methodology is organised into two key stages: 1) The *Concept Detection* stage, where the LVLM predicts the individual presence of each predefined clinical concept in the query image. This is achieved using a custom prompt that supports both zero- and few-shot settings. In the latter, we include a set of demonstration examples (middle block of *Prompt Construction*) chosen by the *Retrieval Module*, responsible for selecting the N most similar examples to the input image. To evaluate the LVLM answer, we employ an *Evaluation Block* which first tries to extract the desired LVLM response using a rule-based formulation. If this fails, we adopt an auxiliary LLM to extract the desired response. 2) In the *Disease Diagnosis* stage, the final diagnosis is generated by the LVLM based on the clinical concepts predicted in the first stage, which are directly incorporated in the query (highlighted in yellow). This approach ensures that the diagnosis is grounded on the identified clinical concepts, enhancing the interpretability and transparency of the LVLM's response. In this second stage, the *Retrieval Module* is also used to select the N most similar demonstrations.

literature in {dermatology, radiology, retinal lesions}, which phrases best describe a {skin, X-ray, eye fundus} image containing {concept}?".

Examples of prompts for different concepts related to the classification of DR are shown in Figure 3.2 on the left. The full prompts for all datasets can be found in Appendix A.

3.3.2 Disease Diagnosis

The second stage involves prompting the LVLM to answer if a certain disease (i.e., class) is present in the image. The concepts predicted in the previous stage are appended to the question, such that when providing an answer, the LVLM not only considers the image and the question, but also takes into account the concepts. Thus, the disease diagnosis is given by:

$$p_{LVLM}(t_{ans}|t_I^1,...,t_I^m,t_P^1,...,t_P^k,t_C^1,...,t_C^c) \tag{3.2}$$

where $t_C^1,...,t_C^c$ are the tokens representing the concepts. This approach ensures that the diagnosis is grounded on the identified clinical concepts, enhancing the interpretability and transparency of

the model's response. The structure of the (zero-shot) prompt for diagnosis classification is as follows:

> {A brief description of {concept 0}}.
>
> (...)
>
> {A brief description of {concept $l-1$}}.
>
> What is the diagnosis that is associated with the following concepts: {predicted concepts}
>
> **Options:**
>
> A) class 0
>
> B) class 1
>
> (...)
>
> Choose one option. Do not provide additional information. Answer:

An example of the prompt designed for the classification of DR is shown in Figure 3.2 on the right. The full prompts for all datasets can be found in Appendix A.

3.3.3 Few-shot Prompting

A key factor contributing to the appeal of LLMs, and consequently LVLMs, is their ability to perform ICL. ICL, or few-shot learning/prompting, is a prompt engineering technique in which the model is provided with examples, known as demonstrations, consisting of: i) questions, and ii) their corresponding answers. It has been shown to significantly improve the performance of both LLMs [131] and LVLMs [132]. Therefore, our methodology incorporates ICL in both stages, as shown in Figure 3.2.

3.3.4 Demonstration Example Retrieval

Although the demonstrations for ICL could be chosen randomly, Zhou *et al.* [133] have shown that this selection policy can decrease the model performance when compared to the 0-shot scenario (when no examples are included in the prompt). Thus, following other works [132], [134]–[136], we introduce a Retrieval Module to select the demonstrations based on their similarity to the query image $\mathbf{x}_0$, a method known as Retrieval-based In-Context Example Selection (RICES). Given the feature vector $\phi(\mathbf{x}_0)$ of the query image obtained from an arbitrary vision encoder, we select the N most similar image feature vectors from the training set $\mathscr{D}$ as the few-shot demonstrations to include in the prompt. For each different type of prompt, the question is repeated for every demonstration. The GT answer to each demonstration question is also provided; depending on the stage of our methodology, this can be the absence/presence of a given concept (1st stage), or the disease diagnosis, as well as the absence/presence of all concepts (2nd stage).

3.3.5 Answer Extraction

As LVLMs produce their outputs in natural language (i.e., it is open-ended), it is necessary to process the generated text to compare it with the GT concepts or the GT class labels. To simplify

this process and limit the space of possible responses, all questions are posed as multiple choice and the LVLM is instructed to answer by choosing one of the options and not providing additional information (see the example prompts in Figure 3.2).

One way of extracting the LVLM answer would be to select the option with the higher first token log probability, i.e., if token B is the one with the highest probability, then this would mean that the LVLM had chosen option B. However, [137] have shown that there is a mismatch between the first token probabilities and the actual LVLM answer. Thus, we use a Regular Expression to directly find the pattern "option_letter)", either in lower or upper case. We choose to provide the options as "option_letter" + ")" instead of "option_letter" + ".", as it would be less probable to find the pattern "option_letter)" in the middle of the LVLM response. If none of the predefined options is found in the LVLM answer (e.g., when the LVLM answers the question but in a more verbose fashion), we follow the approach of previous works [115] and employ an auxiliary LLM[1] to extract the information from the LVLM answer and match it to the predefined list of options. If both approaches fail (e.g., when the LVLM response is just a description of the input image without any actual answer to the question), we mark that instance as unknown and report separately the percentage of instances for which this occurred.

3.4 Experimental Setup

In this section, we outline our experimental setup with information about the adopted LVLMs, datasets, and evaluation metrics, as well as the implementation details.

3.4.1 Datasets

We conduct experiments on four publicly available datasets, encompassing dermatology (Derm7pt [21] and SkinCon [23]), radiology (COVID Radiographic Images Dataset for AI (CORDA) [19]) and eye fundus imaging (DDR [22]) modalities. The selection of these datasets is constrained by the availability of datasets with annotated clinical features, hereinafter referred to as "*concepts*". Unless otherwise stated, we use the official train-val-test splits of the datasets.

For the Derm7pt [21] dataset, with a total of 827 images, we consider only dermoscopic images of the "nevus" and "melanoma" classes, jointly with the 5 annotated dermoscopic attributes, namely Absent / Typical / Atypical Pigment Network, Absent / Regular/ Irregular Streaks, Absent / Regular / Irregular Dots and Globules, Blue-Whitish Veils, and Regression Structures.

The SkinCon [23] dataset contains 3,230 images from the Fitzpatrick 17k dataset [138] annotated with 48 clinical concepts. Following previous works, we use the binary labels denoting skin malignancy for the target task and randomly split the dataset into train/val/test with the proportion of 70%, 15%, and 15%, respectively. Additionally, only 22 out of the 48 concepts are used, as the other 26 concepts appear in less than 50 images.

[1] `Mistral-7B-Instruct-v0.3`

From the CORDA [19] dataset, a COVID-19 diagnosis dataset, we use its 1601 CXRs. Inspired by the work of Barbano *et al.* [139], we use a pretrained model on the CheXpert [140] dataset to annotate the samples with a binary label to denote the presence or absence of 14 radiological observations. Of those 14 concepts, we exclude the "Fracture" and "Support Devices" concepts, as these are not directly related to the presence or absence of the target class. The "No Finding" concept is implicitly included in the remaining 11 concepts given that if all 11 concepts are 0 it represents a healthy case without any findings.

The DDR [22] dataset comprises 13,673 fundus images divided into six DR levels. In our experiments, we focus on the five-class classification task for DR grading, discarding the images belonging to the ungradable class and considering a subset of 757 images annotated with 4 types of lesions correlated with DR, which we use as concepts. We augment this subset with 743 "No DR" images, resulting in a dataset of 1,500 samples.

Table 3.1 provides a detailed summary of the statistics for each of the considered datasets. All results (*cf.* Section 3.5) are reported on the entire test set of each dataset.

Table 3.1: Detailed statistics for the datasets used.

Dataset	Image Type & Task	Classes	Concepts	Train	Test
Derm7pt [21]	Dermoscopic Melanoma classification	2	5	346	320
SkinCon [23]	Dermoscopic Malignancy classification	2	22	2582	554
CORDA [19]	X-ray COVID-19 classification	2	11	967	392
DDR [22]	Fundus DR classification	5	4	817	404

3.4.2 LVLMs

In our experiments, we assess the performance of open-source LVLMs that can process both text and multiple images. Specifically, we focus on models trained on general-domain data, such as OpenFlamingo [134], Idefics3 [141], VILA [142], LLaVA-OneVision [143], Qwen2-VL [144], MiniCPM-V [145], InternVL 2.5 [146] and mPLUG-Owl3 [147]. Additionally, we consider models specialised for the medical domain, including Med-Flamingo [113], LLaVA-Med [114], CheXagent [115] and SkinGPT-4 [116]. We use their corresponding checkpoints from the HuggingFace library (the full list of checkpoints can be found in Appendix A).

3.4.3 Baselines

We compare the performance of *CBVLM* against several baselines, including:

i) standard CBM [125] trained on each of the four datasets considered using the official implementation. It is worth noting that, for training the CBMs on the Derm7pt dataset, we considered 11 clinical concepts. Specifically, 3 out of the 5 annotated concepts have three options each (e.g., Pigment Network can be Absent, Regular, or Irregular), as described in 3.4.1. To address this, and given that the concept layer predicts the presence or absence of each concept, we treated each option as an independent concept. For instance, for Pigment Network, we defined the concepts: Absent Pigment Network, Regular Pigment Network, and Irregular Pigment Network;

ii) Concept-based Lesion Aware Transformer (CLAT) [148] - a concept-based framework specifically designed for retinal disease diagnosis. For DDR, the pre-computed textual embeddings of the clinical concepts provided in the official CLAT repository were used. However, since the original embeddings were obtained from a retinal foundation model [149], for the other datasets, we replaced it with MedImageInsight [150], an embedding model for medical imaging trained on a mixture of X-ray, computed tomography, dermatology, and pathology datasets. For Derm7pt, we follow the same strategy described above, i.e., we transformed the 5 concepts into 11 independent concepts;

iii) supervised black-box models, namely ImageNet-pretrained ResNet50 [89] and ViT Base [47] fine-tuned on the disease diagnosis task;

iv) task-specific black-box models, namely [128] for the Derm7pt dataset, [151] for the SkinCon dataset, [152] for the CORDA dataset, and [153] for the DDR dataset, reporting classification performance on the respective dataset.

3.4.4 Retrieval Module

For the retrieval strategy used to perform ICL, image similarity is determined using cosine similarity between features, which are extracted from two distinct encoders, namely MedImageInsight [150] and the (vision) encoder of the corresponding LVLM.

3.4.5 Evaluation Metrics

For both concept prediction and disease diagnosis, we report Balanced Accuracy (BACC) and F1-score, as these metrics are tailored to deal with imbalanced datasets, which is the case in medical image classification. For concept prediction, we measure the BACC and F1-score averaged over each concept.

3.5 Results and Analysis

This section presents the key results of *CBVLM*, comparing its performance against CBMs and task-specific supervised methods in terms of BACC and F1-score across two tasks: i) Concept Detection (Figures 3.3, 3.4, 3.5, and 3.6); and ii) Disease Diagnosis (Figures 3.7 and 3.8).

3.5.1 Concept Detection Performance

We start by assessing the predictive performance of *CBVLM* in the task of concept detection. Figure 3.3 presents the results across different datasets under various n-shot settings, $n \in \{0,1,2,4\}$. For comparative analysis, we also report the performance of CBM and CLAT. The results for $n > 0$ were obtained considering the best vision encoder, in this case, MedImageInsight, selected after an ablation study with different vision encoders (see Figure 3.9).

3.5.1.1 Few-shot Prompting Boosts Performance

The results presented in Figure 3.3 and the analysis depicted in Figure 3.4 (left) reveals that few-shot prompting significantly boosts performance by incorporating contextually relevant information within the prompts [113]. Notably, there is a consistent trend across all datasets, demonstrating improved performance as the number of demonstrations increases. For example, the performance of *CBVLM* when adopting CheXagent, LLaVA-OV, Qwen2-VL, and SkinGPT-4 improves by over 15% when incorporating 4 demonstration examples into the prompt compared to the zero-shot scenario ($n = 0$). Remarkably, *CBVLM* outperforms both CBM and CLAT across all datasets except Derm7pt, where its performance is on par with CBM . Impressively, *CBVLM* achieves these results without any training, relying on only four demonstration examples or less.

3.5.1.2 Medical LVLMs Outperform Generic LVLMs in Few-shot Settings

From the results depicted in Figure 3.4 (right), in the 0-shot scenario, generic LVLMs consistently outperform medical LVLMs across all datasets. However, in the few-shot setting, medical LVLMs outperform their generic counterparts in all datasets using only one or two demonstration examples ($n \in \{1,2\}$). The stronger performance of generic LVLMs in the zero-shot setting can be attributed to their training on large-scale generic image datasets, allowing for better generalization in this setting. As demonstration examples are introduced into the prompt, medical LVLMs improve, benefiting from their pre-training on specialised medical data, which enhances their ability to generalize in these scenarios. However, in the 4-shot setting, generic LVLMs tend to achiever higher concept detection performance than their medical counterparts. This might stem from the fact that generic LVLMs tend to allow for longer prompts and are trained with interleaved image-text data, while medical LVLMs are not.

3.5.1.3 Underrepresented Concepts are Harder to Predict

Figure 3.5 presents a detailed comparison of the per-concept F1-score for each dataset. The reported results, averaged across all, generic, and medical LVLMs in the 4-shot setting, reveal that underrepresented concepts are the most challenging to detect, as expected. For instance, in the CORDA dataset, none of the models correctly predicts the presence of the "pleural other" concept (which only appears in 4 images of the entire dataset). The concepts "atelectasis" and "lung lesion" are also consistently difficult to detect. The full results per-LVLM can be found in Appendix A.



Figure 3.3: Concept detection performance of LVLMs across different n-shot settings. Each bar corresponds to an n-shot scenario ($n = \{0,1,2,4\}$). Filled coloured bars denote BACC, whereas the hatched bars indicate F1-scores. Red crosses indicate the percentage of unknowns, i.e., the proportion of samples whose LVLMs' responses do not contain sufficient information to answer the posed question.

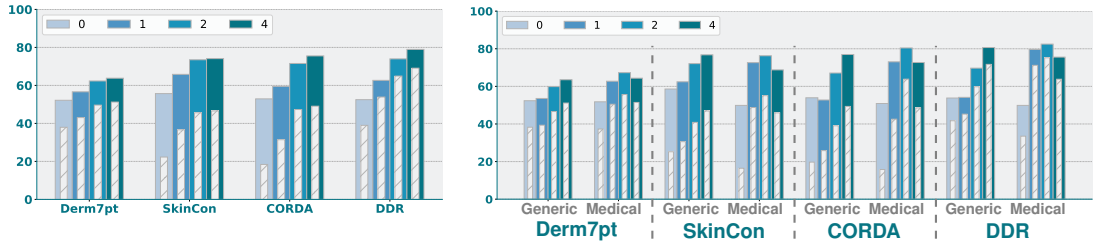


Figure 3.4: Concept detection results per dataset averaged over all models (left) and over generic and medical LVLMs (right). Each bar corresponds to the number of shots ($n = \{0,1,2,4\}$). Filled coloured bars denote BACC, whereas the hatched bars indicate F1-scores.

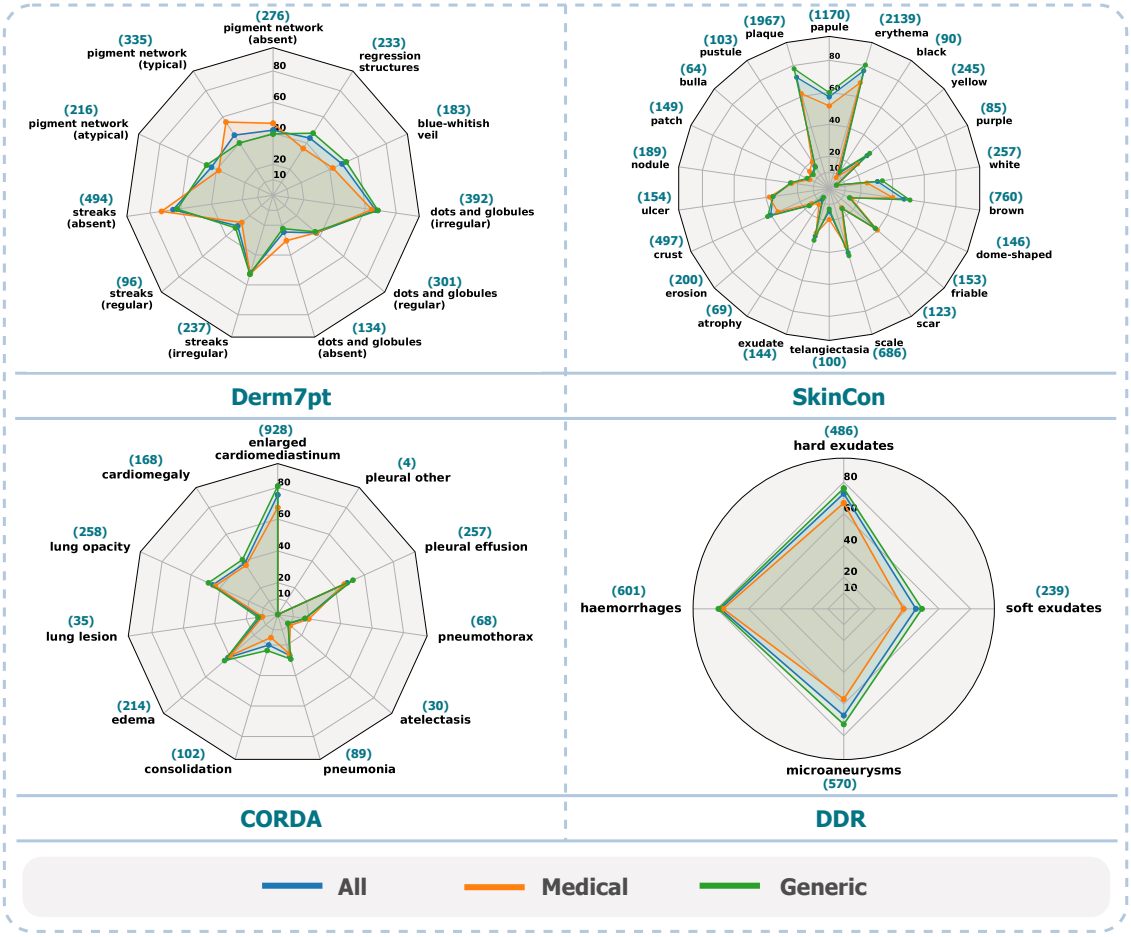


Figure 3.5: Per-concept F1-scores across datasets averaged over all (blue), generic (green), and medical (orange) LVLMs, in the 4-shot scenario. Values in parentheses indicate the number of samples in the dataset for the corresponding concept.

3.5.1.4 *CBVLM* Outperforms CBM

CBVLM outperforms CBM and CLAT in concept prediction across all datasets, except on Derm7pt, where CBM achieves the best result with a margin of $\approx 5\%$. This may be due to the fact that in Derm7pt, some concepts have three available options, whereas in the other datasets, each concept is binary, which may make it easier for the LVLM to predict the presence or absence of a concept more accurately.

3.5.1.5 *CBVLM* Outperforms CBM When Using Only 10% Annotated Data

Although *CBVLM* is able to outperform CBM with only 4-shots, it should be noted that, in total, more than 4 images are being used, as the same 4 example images will not be the most similar for all query images in the test set. To assess the impact of the number of images per class in the examples set, we progressively reduce the percentage of images per class, limiting the number of available examples for selection for few-shot prompting. This experiment is conducted using

the best-performing model for each dataset: CheXagent 4-shots for Derm7pt, and Open-Flamingo 2-shots for SkinCon, CORDA and DDR. The results, shown in Figure 3.6 and averaged across 3 runs, demonstrate that even with just 10% of the images per class in the examples set, *CBVLM* consistently outperforms CBM across all datasets, except for Derm7pt where even with 100% of the images per class, the performance is lower than CBM , as discussed earlier. Nevertheless, the results shown in Figure 3.6 underscore a key observation: few-shot prompting with a good example selection strategy can boost performance by up to 20% (*cf.* BACC at 0% and 10%), even in situations where annotated data is severely limited, which is the case of medical data. Thus, LVLMs can indeed be successfully used for concept detection with very limited amounts of annotated data, thus reducing the annotation cost inherent to CBMs.

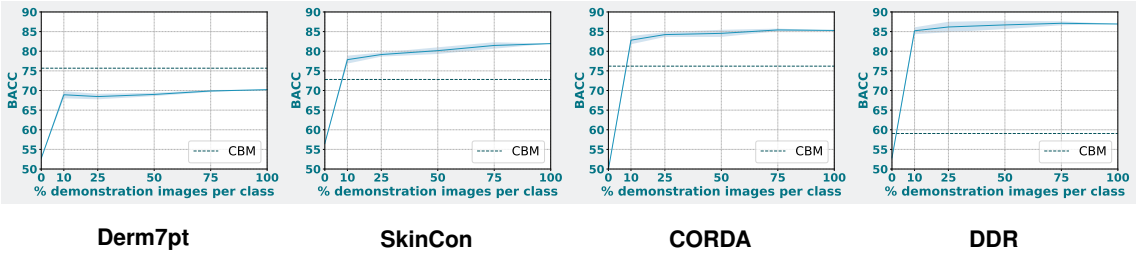


Figure 3.6: Influence of the example set size on the concept prediction performance of *CBVLM*. Even when only 10% of the example images per class are available, *CBVLM* outperforms CBM, except for Derm7pt. Nevertheless, with only 10% of the examples, the decrease in performance is negligible for Derm7pt. Thus, *CBVLM* indeed requires few annotated examples.

3.5.2 Disease Diagnosis Performance

Figure 3.7 shows the performance of *CBVLM* in terms of BACC and F1-score for the task of disease classification across different datasets under various n-shots ($n = \{0, 1, 2, 4\}$). Similarly to Section 3.5.1, we also report the performance of CBM , CLAT and a Black-box (Bbox) model for comparison. Additionally, we include results for the setting where the *CBVLM* predicts the disease label directly without prior concept information ("0 w/o" bar in Figure 3.7).

3.5.2.1 Few-shot Prompting Boosts Performance

Similar to the trend observed in the concept detection task, the results shown in Figure 3.8 (left) confirm our initial assumption and align with findings reported in the literature: few-shot prompting significantly enhances the performance of LVLMs in disease classification as the number of demonstrations increases, across all datasets.

3.5.2.2 Using Concepts Improves Performance and Promotes Explainability

The results show that incorporating clinical concepts into the prompt improves both BACC and F1-score compared to the scenario without concept guidance (i.e., "0" vs "0 w/o" bars in Figure 3.7).



Figure 3.7: Disease diagnosis performance of LVLMs across different n-shot settings. Each bar corresponds to an n-shot scenario ($n = \{0, 1, 2, 4\}$). "0 w/o" corresponds to the 0-shot experiment, where no concepts are used for the disease diagnosis. Filled coloured bars denote BACC whereas the hatched bars indicate F1-scores. Red crosses indicate the percentage of unknowns, i.e., the proportion of samples whose LVLMs responses do not contain sufficient information to answer the posed question.

Furthermore, grounding the final diagnosis on a set of clinical concepts promotes transparency and explainability, aligning more closely with human reasoning, unlike the concept-free setting ("0 w/o").

3.5.2.3 Medical LVLMs Outperform Generic LVLMs in Few-shot Settings

The results in Figure 3.8 (right) confirm that generic LVLMs consistently outperform their medical counterparts across all datasets in the scenario without concepts ("0 w/o"). This trend is also observed in the 0-shot scenario with concepts, in general. In the few-shot setting, generic LVLMs exhibit superior performance only in SkinCon in terms of BACC, with a marginal average improvement of 1% across scenarios with $n > 0$. Conversely, when averaging across all

few-shot scenarios ($n > 0$), medical LVLMs consistently outperform generic models across all datasets, achieving F1-score improvements of 25.74%, 14.48%, 4.62%, and 19.81% for Derm7pt, SkinCon, CORDA, and DDR, respectively. This phenomenon can be attributed to two factors: i) the stronger domain knowledge of medical LVLMs, and ii) the initialization of their vision encoders from medical-specific Contrastive Language-Image Pre-training (CLIP) versions, such as BiomedCLIP [154] for LLaVA-Med, or larger improved CLIP architectures like EVA-CLIP-g [155], used in CheXagent and SkinGPT.

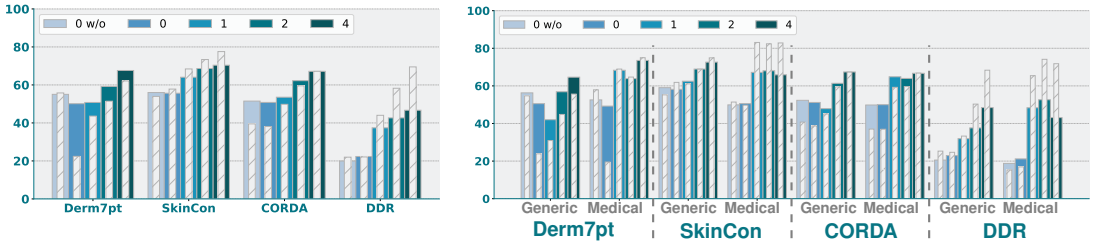


Figure 3.8: Disease diagnosis results per dataset averaged over all models (left) and over generic and medical LVLMs (right). Each bar corresponds to the number of shots ($n = \{0, 1, 2, 4\}$). "0 w/o" corresponds to the 0-shot experiment where no concepts are used for the disease diagnosis. Filled coloured bars denote BACC, whereas the hatched bars indicate F1-scores.

3.5.2.4 *CBVLM* Outperforms CBM

CBVLM surpasses CBMs across all datasets. With the top-performing LVLM, we observe F1-score improvements of 7.42%, 39.70%, 7.25%, and 63.13% for Derm7pt, SkinCon, CORDA, and DDR, respectively, compared to the standard CBM . It is worth recalling that *CBVLM* achieves these results without any kind of training, whereas CBM needs to be trained for each dataset. When comparing *CBVLM* with CLAT [148], a method specifically optimised for retinal disease diagnosis, our experiments show that *CBVLM* outperforms it on the DR dataset, DDR, in terms of F1-score. Furthermore, *CBVLM* also outperforms CLAT on the other datasets, achieving an average relative improvement of $\approx 6\%$.

3.5.2.5 *CBVLM* Outperforms Supervised Black-box Models

Finally, when compared to black-box approaches, a similar trend is observed. *CBVLM* outperforms task-specific black-box models by 2.15% in BACC for Derm7pt and achieves comparable performance on CORDA, with a slight decrease of -2.32%. In terms of F1 score, the best performing *CBVLM* exceeds black-box task-specific models by 0.36% on SkinCon and by 1% on DDR. Thus, *CBVLM* achieves state-of-the-art results on Derm7pt, SkinCon and DDR.

3.5.3 Ablation Study: Strategies for ICL Example Selection

In order to validate the importance of the *Retrieval Module* (*cf.* Figure 3.2), which employs RICES [135] (i.e., ranks examples based on the cosine similarity between image features extracted

from an arbitrary vision encoder), we conduct an ablation study on the task of concept detection for the 1-shot scenario.

In this study, we first compare random ICL example selection with RICES, and then assess different vision encoders to perform RICES. For random selection, we compare choosing only 1 random example against choosing 1 random example per class. The results of Figure 3.9 show that, except for Open-Flamingo and MiniCPM-V, using 1 example per class does not improve performance. In fact, in the majority of cases it slightly decreases it, when comparing to simply choose 1 random example.

We then explore two different encoders for RICES: MedImageInsight [150] and the encoder integrated within the respective LVLM. Although in some models (e.g., Open-Flamingo and MiniCPM-V), choosing the examples randomly brings better performance, in the majority of cases it improves both BACC and F1-score (especially for Medical LVLMs). More importantly, we are not concerned with the average results, but rather on maximising them, even if just for one LVLM, to be able to find the best configuration of *CBVLM* to compete against other approaches like CBMs and supervised black-box models. Thus, using RICES beats random selection and using MedImageInsight instead of the LVLM encoder yields the best results on all datasets. As such, all experiments are conducted using RICES (i.e., the *Retrieval Module*) with MedImageInsight as the feature extractor.

3.5.4 General Discussion

Table 3.2 presents a detailed comparison of our proposed *CBVLM* with concept-based models, supervised black-box models, and task-specific models. As highlighted in the results, *CBVLM* consistently achieves either the best or second-best performance across all evaluations. Notably, this is accomplished without any additional training and by leveraging just four examples, enabling concept-based explanations and enhanced interpretability and transparency in the model responses. In summary, *CBVLM* using CheXagent with 4 demonstration examples emerges as the top-performing combination for Derm7pt. For SkinCon, *CBVLM* with the generic Open-Flamingo model and 4 demonstration examples achieves the best results in both disease diagnosis and concept detection tasks. For CORDA, the best performance is achieved by *CBVLM* with CheXagent with 2 demonstration examples. For DDR, *CBVLM* with Idefics3 and 4 demonstrations proves to be the best-performing combination.

Overall, the results underscore the effectiveness of *CBVLM* in enabling arbitrary LVLMs to predict the presence of clinical concepts from an image and use those concepts to ground the final diagnosis, enhancing interpretability and transparency in the model's responses, all without requiring any training and with just a few annotated examples. Notably, *CBVLM* outperforms CBM -related approaches across all datasets and even surpasses black-box models when employing ICL.



Figure 3.9: Concept detection results with 1-shot and several example selection strategies. "random" corresponds to random selection of 1 example, while "random_per_class" corresponds to selecting 1 example per class randomly. "rices" is related to the RICES methodology [135], where the demonstration example is chosen based on the cosine similarity to the query image's features. "rices_medii" corresponds to using RICES when MedImageInsight [150] is the vision encoder used to extract the features from the query and demonstration images, while in "rices_model" the encoder is the vision model included in the corresponding LVLM.

3.5.5 Limitations

The main drawback of the proposed methodology for generating concept-based explanations with LVLMs lies in the independent prediction of each concept, which is a concern when applied to datasets with a large number of concepts. A similar problem arises with datasets with a large number of classes, as each class is provided as a separate option in the prompt. This can result in excessively long prompts that would require alternative ways of querying the model. Although in medical applications the number of concepts is typically limited, a potential future direction to address concerns regarding the scalability for datasets with numerous concepts could be the adoption of tailored selection policies that prioritize only a small subset of the most informative

Table 3.2: Comparison of the best performing *CBVLM* with state-of-the-art supervised models and CBM for disease diagnosis and concept detection. BACC refers to Balanced Accuracy, and F1 refers to the F1-score. The best performance is highlighted in **bold**, while the second-best performance is indicated with underlining.

Dataset	Method	Disease Diagnosis		Concept Detection	
		BACC	F1	BACC	F1
Derm7pt	CBM [125]	74.01	75.66	**75.68**	**67.11**
	CLAT [148]	79.67	82.98	66.69	54.76
	Black-box (ResNet50)	75.33	80.76	-	-
	Black-box (ViT Base)	82.26	82.04	-	-
	Black-box (Task-Specific) [128]	83.20	-	-	-
	CBVLM (CheXagent & 4-shot)	**85.35**	**83.08**	70.21	59.55
SkinCon	CBM [125]	49.97	46.51	72.81	56.33
	CLAT [148]	68.21	**86.76**	62.64	45.52
	Black-box (ResNet50)	73.98	81.49	-	-
	Black-box (ViT Base)	80.25	85.85	-	-
	Black-box (Task-Specific) [151]	-	77.80	-	-
	CBVLM (Open-Flamingo & 4-shot)	**85.61**	86.21	**79.78**	**57.91**
CORDA	CBM [125]	70.75	70.98	76.20	61.92
	CLAT [148]	75.64	77.96	70.76	53.06
	Black-box (ResNet50)	73.17	73.28	-	-
	Black-box (ViT Base)	78.03	77.63	-	-
	Black-box (Task-Specific) [152]	**80.00**	-	-	-
	CBVLM (CheXagent & 2-shot)	77.68	**78.23**	**81.38**	**64.33**
DDR	CBM [125]	23.62	21.25	59.05	39.06
	CLAT [148]	**78.87**	72.81	76.64	64.53
	Black-box (ResNet50)	33.70	57.00	-	-
	Black-box (ViT Base)	39.09	59.37	-	-
	Black-box (Task-Specific) [153]	-	83.38	-	-
	CBVLM (Idefics3 & 4-shot)	66.65	**84.38**	**82.69**	**75.73**

concepts, as demonstrated in recent works [156].

Additionally, our approach is limited to discrete-valued concepts (e.g., absent/present or, for some concepts of Derm7pt, absent/regular/irregular), contrasting with traditional CBMs, where each concept is associated with a probability. Nevertheless, this limitation could be addressed by prompting the LVLM to not only answer if a given concept is present in the input image but also to provide a confidence score, as is done in other works [119]. We leave this exploration to future work.

Finally, *CBVLM* is only as good as the LVLM it uses. In a few cases, some of the LVLMs we tested struggled to provide answers to the questions (e.g., instead of answering the question, the LVLM simply described the input image), leading to an increase in unknown responses (as shown in Figures 3.3 and 3.7). This phenomenon is particularly common with SkinGPT, which may be

due to its pretraining on very specific prompts designed to elicit desired responses [116]. LLava-Med also faced challenges in providing valid responses, especially in few-shot scenarios where $n > 1$. This difficulty may stem from the fact that it is not trained with interleaved image-text data; in such models, images are given at the input in the form of tokens, so as the number of shots increases, so does the number of tokens, making it more difficult for the model to pay attention to the full input sequence [157].

Another relevant concern involves the robustness, security, and inherent biases of LVLMs [158]. While a comprehensive analysis of these aspects is beyond the scope of this work, it remains a critical consideration for future studies aimed at safe and reliable integration of LVLMs into clinical workflows. Collaboration between clinicians and AI researchers is crucial to ensure transparency and explainability of AI models deployed into clinical practice [159].

3.6 Summary and Conclusions

In this chapter, we proposed a simple yet effective methodology, *CBVLM*, that leverages off-the-shelf LVLMs to produce concept-based explanations and predict disease diagnoses grounded on those explanations. Through extensive experiments across four medical datasets, and open-source LVLMs, we showed that *CBVLM* outperforms the traditional CBMs and even task-specific black-box models across all benchmarks. This is achieved without requiring any training and using just a few annotated examples, thus fulfilling the two main goals of this thesis: enhancing interpretability without increasing the annotation burden. Overall, *CBVLM* is able to leverage the best of both worlds: (i) its LVLM-based training-free nature presents a major advantage over CBMs, as these models require additional training and are restricted by predefined concepts, and (ii) its two-stage process keeps the interpretability inherent to CBMs but that LVLMs lack.

Chapter 4

Natural Language Explanations

Foreword on Author Contributions

The research work described in this chapter was conducted under the supervision of Professors Jaime S. Cardoso and Luís F. Teixeira.

The results of these works have been disseminated in three articles, two in international conferences and one under revision in an international journal:

[25] Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "From Captions to Explanations: A Multimodal Transformer-based Architecture for Natural Language Explanation Generation" Pattern Recognition and Image Analysis, pp. 54–65, 2022

[26] Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "Parameter-Efficient Generation of Natural Language Explanations for Chest X-ray Classification" Medical Imaging with Deep Learning (MIDL), vol. 250, pp. 1267–1281, 2024

Code: https://github.com/icrto/peft-nles

[27] Isabel Rio-Torto, Jaime S. Cardoso, and Luís F. Teixeira, "WeNLEX: Weakly Supervised Natural Language Explanations for Multilabel Chest X-ray Classification" under revision at IEEE Transactions on Medical Imaging (TMI)

Code: https://github.com/icrto/WeNLEX

Another type of explanations which are inherently interpretable to humans are NLEs, the topic of this chapter. Naturally, NLEs can complement both modalities explored in previous chapters. For example, unlike visual explanations, which are spatially precise but often semantically limited (and can be hard to understand [100]), NLEs are fundamentally human-understandable, closely reflecting the way radiologists convey their findings in textual reports [160], [161].

We begin by addressing the problem in a fully supervised setting (Section 4.5) to demonstrate that it is feasible to supervise NLE generation at the Encoder's output within an Encoder–Decoder architecture. After empirically validating this approach, we move to our ultimate objective: gen-

erating faithful and plausible NLEs in a weakly supervised setting through the introduction of a novel model called *WeNLEX* (Section 4.6).

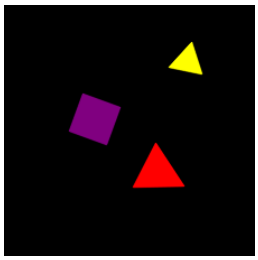
4.1 Context and Motivation

NLEs are textual descriptions that go beyond image captioning (image-relevance), since besides the input image they also need to take into account the underlying model's decision-making process [10], [162]. The same way a visual explanation differs from a segmentation map [163], so does a textual explanation differ from a caption [164]. While the latter constitutes a detailed description of an image, ideally mentioning all the existing objects, an explanation should only refer to (part of) objects that are relevant for the model decision.

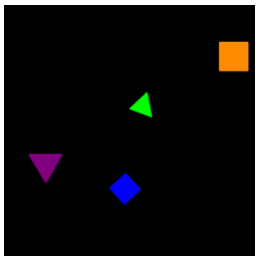
In traditional image captioning datasets, such as Common Objects in Context (COCO) [30], captions tend to overlap with explanations because images usually have a single object present and classification is based on the presence of that object; for example a caption could be "An image of a dog laying down." while the explanation could be "There is a dog in the image.". Some efforts have been made to collect human-annotated explanations for existing datasets, but these are either for purely textual data [165], activity recognition [166], Visual Question Answering (VQA) [166], or visual entailment [167]. Therefore, there is a clear lack of image classification data with associated NLEs.

An example where there is an evident distinction between an image caption and an NLE is shown in Figure 4.1. In this synthetic classification scenario, an image is labeled as positive if there is at least one triangle on the left of the leftmost square, and is considered negative otherwise (if no squares/triangles exist or the leftmost figure is not a triangle). As such, there is an unmistakable contrast between a caption, which ideally should account for all the polygons and their relative positions, and an explanation for the classification label, which should only need to identify the presence of the leftmost square and focus on the objects on its left.

Current NLE generation approaches face critical limitations. Most approach the problem as a supervised image captioning task, where the explanation model learns to produce human-annotated explanations. This not only imposes a significant annotation burden (requiring fully annotated explanation datasets with one NLE per image per label) but also assumes that the reasoning of the model being explained (MBE) is similar to the reasoning of the human annotators. In other words, fully supervised NLEs reflect the annotations (are plausible) but not the model's true reasoning (are not faithful). This explicit supervision also makes it impossible to fully explain pretrained models (i.e., produce *post-hoc* explanations): we would need to have GT NLEs for each combination of model predictions (either correct or incorrect), and this set of GT NLEs would, of course, differ from model to model. In fact, being able to explain incorrect model decisions is crucial, e.g., for debugging purposes, and it constitutes a clear case where the model reasoning does not align with the human's, hence the model prediction being defined as incorrect. Therefore, our main focus lies in generating NLEs without direct supervision. However, this implies that there is no GT explanation during training, thus rendering it impossible to train autoregressive



Label: negative
Caption: There are a total of 3 shapes, of which 1 is a square and of which 2 are triangles. The square is purple. One triangle is red and the other is yellow. The polygon on the right is the yellow triangle and the polygon on the left is the purple square. The figure in the middle is the red triangle.
Explanation: The purple square is the leftmost figure.



Label: positive
Caption: Four geometric figures, 2 squares and 2 triangles. One triangle is purple and another is green. One square is orange and another is blue. The rightmost polygon is the orange square and the leftmost polygon is the purple triangle. The green triangle and the blue square are in the middle of the orange square and the purple triangle. The green triangle is on the right of the blue square.
Explanation: The purple triangle is on the left of the blue square.

Figure 4.1: Synthetic examples clearly evidencing the difference between image caption and explanation.

language models in their most common and efficient way, i.e., using teacher forcing. One possible approach would be to use Reinforcement Learning (RL), but this would make training slower and potentially less stable.

As a first stepping stone towards using less supervision in NLE generation, we start by proposing replacing the NLE Decoder-only generator with a denoising Autoencoder (AE) (see Section 4.5). By doing this, the learning can now happen on the continuous latent space of the encoder at the sentence level, and gradient propagation is ensured without needing RL. It also speeds up the training process since obtaining an NLE now only involves a single encoder forward pass instead of N decoder passes, where N is the number of tokens in the generated sentence. Lastly, decoding this latent space into text to obtain the NLE is still possible thanks to the pretrained decoder. We empirically demonstrate, on a multilabel CXR classification task, that this approach is viable.

Therefore, in Section 4.6, we build upon the previous work and move to the weakly supervised generation of NLEs, with the proposal of a novel model named *WeNLEX*. *WeNLEX* ensures faithful NLEs by matching (in the MBE's feature space) the images generated from the NLEs with the original input images. Plausibility is maintained through distribution alignment with a small set of clinician-annotated NLEs. We show that *WeNLEX*: i) generates faithful NLEs, requiring as few as five GT explanations per diagnosis; ii) can operate in both *post-hoc* and in-model settings, even improving classification performance by 2.21% Area Under the Curve (AUC), demonstrating that interpretability can enhance task performance; and iii) offers adaptability to different target audiences, exemplified by a layman version that produces simplified explanations for non-expert users.

4.2 Related Work

The literature on the generation of NLEs for text-only [168]–[171], and for image-only or image-text [166], [172]–[178] tasks is vast. Consequently, there are many architectures for NLE generation, which can be grouped according to different criteria:

- network architecture: the type of neural network, e.g., CNNs [166], [172], [173], [179], [180] vs. vision transformers [174]–[179], [181] for image processing, and RNNs [166], [172] vs. transformer-based models [174]–[176], [179]–[181] for text generation
- modularity: separation between task prediction and explanation, e.g., modular (separate modules for task prediction and explanation generation) [166], [172], [174]–[176], [179], [180] vs. unified (single integrated model) [177], [178], [181]
- training paradigm: whether the task module is trained separately or jointly with the explanation model, e.g., *post-hoc* (task model pretrained and frozen) [166], [172], [174], [175], [179], [180] vs. in-model (task model and explanation generator trained together) [176]–[178], [181]
- reasoning flow: the order in which prediction and explanation are generated, e.g., *predict-explain* (prediction first, then explanation) [166], [172], [174], [175], [177]–[181] vs. *explain-predict* (prediction is based on the explanation) [176]

In the medical domain, to the best of our knowledge, there are only two works on NLE generation [24], [182]. Kayser *et al.* [24] introduce the MIMIC-NLE dataset and test different architectures (Radiological Text Captioning for Human Examined Thoraces (RATCHET) [183] and DenseNet [88] + GPT-2 [184] (DPT)) to generate NLEs for multilabel CXR classification. All have the same working principles: i) a vision model extracts features and classifies the input image, and ii) for each predicted diagnosis, the features, the whole multilabel prediction vector, and the pathology for which the NLE is being generated, are given to a language model, which autoregressively generates the NLE.

More recently, Hamza *et al.* [182] propose a Knowledge Graph (KG) Retrieval-Augmented Generation (RAG) framework to enhance the Low-Rank Adaptation (LoRA)-based [185] fine-tuning of the LLaVA [186] vision-language model. For effective knowledge retrieval, the authors build a knowledge graph of relationships between medical entities extracted from the medical reports of MIMIC-CXR [13] using the RadGraph model [187]. For each image, the closest KG entities according to MedCLIP [188] are retrieved and given to the language model of LLaVa [186], together with the projected image embeddings and the vision classifier. By incorporating domain-specific knowledge, this approach achieves state-of-the-art results on MIMIC-NLE.

Despite the variety of NLE generation models, all aforementioned works fully supervise the NLE generation process by computing the cross entropy between generated and human annotated NLEs. As previously argued, this renders their NLEs unfaithful, only reflecting the annotations, instead of the underlying model's decision process.

To the best of our knowledge, there are only two works that tackle NLE generation without direct supervision [10], [164]. The first was among the initial works on NLE generation and is still perhaps one of the most relevant. It certainly has inspired the work we present in this chapter. Hendricks *et al.* introduce an architecture composed of a compact bilinear classifier for image feature extraction and a RNN conditioned on the predicted class label. This work distinguishes itself because NLEs are obtained by using the supervision of GT image captions conditioned on the predicted label, through a RL-based discriminative loss that rewards explanations with class discriminative properties. The rewards are given by a sentence classifier trained to predict the class label from an input text.

The other is the work of Sammani and Deligiannis [10], which tackles the problem of generating NLEs for vision classification in a zero-shot manner. Their method starts by training a simple Multilayer Perceptron (MLP) on top of a textual encoder to predict the vision classifier's weights for each class in the dataset. With this MLP, any text can be mapped into the vision classifier embedding space, and the distance between text and visual features can be measured in this shared space. Then, at inference time, learnable prefixes are optimised to steer an off-the-shelf language model already capable of generating plausible text, towards generating an NLE that maximizes the similarity with visual features through the text encoder+MLP. Regardless of being a plug-and-play approach, easily adaptable to any vision classifier, it would not be applicable in the case of MIMIC-NLE for two reasons: i) the MLP is originally trained for the 1k classes of ImageNet [31] and it is not clear if it would be able to achieve reasonable performance for only 10 classes, and ii) it cannot operate in a multilabel setting.

4.3 Dataset

In the general domain, image captioning datasets abound [28], [30], [189]. Datasets with NLEs are more scarce, but do exist, both for text-only [168], [169], [190], [191] and vision-language tasks [166], [173], [175]. Wiegreffe and Marasović [192] provide a comprehensive review on the topic.

In the medical domain the same trend is verified: there are some datasets that include medical reports [13], [193]–[195], but datasets with NLEs are rarer. In fact, to the best of our knowledge, MIMIC-NLE [24] is the only dataset with NLEs for CXR classification. Hence, this is the dataset that is used in all experiments detailed in the remainder of this chapter.

The MIMIC-NLE dataset is extracted automatically from the medical reports of MIMIC-CXR [13] through clinically validated extraction rules together with the CheXbert labeler [196]. It comprises 38003 image-NLE pairs or 44935 image-label-NLE triplets (one NLE can explain multiple diagnoses/labels). One image can be associated with 10 ($L = 10$) pathologies simultaneously, each with 3 possibilities ($C = 3$): *Positive* (there is clear evidence of the presence of that pathology in the image), *Uncertain* (the pathology might be present in the image), and *Negative* (there is clear evidence of the absence of that pathology in the image). The 10 pathologies are divided

into: i) diagnoses labels, i.e., the labels being explained by the NLEs (*Atelectasis*, *Consolidation*, *Edema*, *Pleural Effusion*, *Pleural Other*, *Pneumonia*, and *Pneumothorax*), and ii) evidence labels, i.e., labels that are part of the evidence that some diagnosis label is present in the input image (*Consolidation*, *Enlarged Cardiomediastinum*, *Lung Lesion*, and *Lung Opacity*). *Consolidation* can be considered both a diagnosis or an evidence label, depending on the interaction with other pathologies (e.g., an image can have only consolidation as a diagnosis label or it can have consolidation as evidence for the existence of pneumonia).

Besides the big imbalance of the *Lung Opacity* evidence label (around 77% of all NLEs), another limitation of the MIMIC-NLE dataset stems from the fact that it uses the CheXbert labeler [196] to extract those evidence labels from NLEs. This means that around 15% of the NLEs do not have any evidence label; this may be because there are actually no evidence keywords in the original NLE (see examples A and B below) or simply that the evidence present in the NLE is not among the 14 labels that CheXbert is able to predict (see examples C and D below).

A. Findings most consistent with moderate pulmonary edema.

B. Findings suggesting mild pulmonary vascular congestion.

C. There is again a coarse reticular abnormality favoring the bases and peripheral aspects of the lung, most consistent with pulmonary fibrosis.

D. There is mild vascular congestion consistent with fluid overload.

In such cases it is not possible to evaluate the evidence present in a generated NLE: even if completely incorrect or devoid of useful information (like examples A and B), any NLE that triggers the CheXbert labeler to output only zeros, would be considered valid.

In the first work that will be further described in this chapter, since we still operate under the fully supervised setting for NLE generation (Section 4.5), we use the original dataset without any modifications (37016, 273, and 714 NLEs for training, validation and testing, respectively). In the second work (Section 4.6), since only a few NLEs from the dataset are used (i.e., it is a weakly supervised scenario) we chose to remove the instances without any evidence labels. So, in Section 4.6, all experiments are carried out with this version of the MIMIC-NLE dataset (35600, 254, and 647 NLEs for training, validation and testing, respectively).

4.4 Problem Formulation

Let us consider a multiclass multilabel classification task, where the goal is to predict a target $y \in \{0,1,...,C\}^L$, where L is the number of labels (of which more than one can be present simultaneously) and C is the number of classes per label, from an input image $x \in \mathbb{R}^{c \times h \times w}$, where c,h,w are the number of channels, height and width, respectively. Let us also consider a classification model, f_{θ_1}, that learns to predict y from x, i.e., $\hat{y} = f_{\theta_1}(x)$. Our goal is to design a system, *WeNLEX*, that generates explanations in natural language, $\hat{e}$, for the predictions of classifier f, such that, $\hat{e} = g_{\theta_2}(x, \hat{y}, f_{\theta_1})$. Thus, it constitutes a predict-explain paradigm. It worth noting that, while in a multiclass task an explain-predict paradigm would be feasible, in a multilabel task such as this

one, it is not: since we generate one NLE per predicted label, we need to know what the prediction is beforehand.

During training of both proposed approaches (Sections 4.5 and 4.6) two scenarios are possible: i) f_{θ_1} has been trained and remains frozen (*post-hoc* scenario), or ii) f_{θ_1} is being trained simultaneously with g_{θ_2} (in-model scenario). In the fully supervised approach (Section 4.5) we only consider the in-model setting, for comparison with state-of-the-art models. In the weakly supervised proposal (Section 4.6) both scenarios are explored.

4.5 Fully Supervised Approach

4.5.1 Baselines

Besides introducing MIMIC-NLE, Kayser *et al.* [24] also tested different architectures to generate NLEs for this multilabel scenario. These architectures, which we describe below, are used here as baselines to compare our proposals with. All architectures work in a similar fashion: a vision model (DenseNet-121 [88]) extracts features and classifies the input image; the features and a vector representation of the prediction, together with the pathology for which the NLE is being generated, are fed into an autoregressive language model, which generates the NLE. During training, instead of using the prediction vector, the GT label is used, and the vision model is trainable. Thus, this can be considered an in-model explanation architecture that works within a predict-explain paradigm [168].

RATCHET was first proposed by Hou *et al.* [183] as a CXR captioning method. Kayser *et al.* [24] leverage this architecture to generate NLEs instead. It consists of a GPT-2 [184] model with cross-attention layers, that receive the vision classifier's features concatenated with a projection of the prediction vector, as depicted in Figure 4.3 (top left). As input, the model receives the NLE-related pathology.

DPT is proposed by Kayser *et al.* [24] and it consists of a GPT-2 model that receives as input the vision features, the prediction vector, and the NLE-related pathology. An illustration of this architecture can be found in Figure 4.3 (top right).

4.5.2 Proposed Approach

4.5.2.1 Architecture

Our approach builds upon the work of Kayser *et al.* [24], but replaces the NLE decoder with a denoising AE (see Figure 4.2). Not only does this increase the representation capabilities of the NLE generator, but, most importantly, it lays the foundation for the RL-free adversarial generation of NLEs when no (or few) GT NLEs are available for training. In this scenario, teacher forcing is unfeasible. While an adversarial approach can solve the problem, if the generator is a decoder-only architecture (e.g., DPT), RL is needed, since its output (and the input to the discriminator) is discrete and, as such, no gradient would flow from the discriminator to the generator [197]. By using an AE architecture, where the adversarial learning happens on the continuous latent space

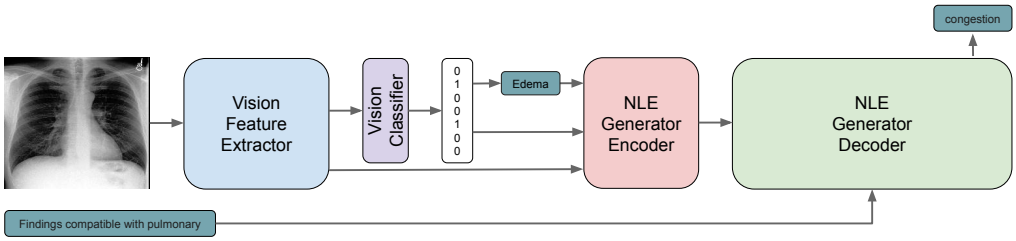


Figure 4.2: Block diagram of the fully supervised NLE generation architecture for the pathology "Edema". The encoder receives the image features and the prediction information, while the decoder generates the NLE from the latent space of the encoder.

of the encoder at the sentence level, gradient propagation is ensured without needing RL, thus making the training process faster (obtaining an NLE now only involves a single encoder forward pass instead of N decoder passes, where N is the number of tokens in the generated sentence) and more stable. Finally, decoding this latent space into text to obtain the NLE is still possible thanks to the pretrained decoder.

We use the AE proposed by [198], which consists of a frozen Bidirectional Encoder Representations from Transformers (BERT)-based encoder, a multi-head attention mechanism as a bottleneck layer, and a single transformer decoder layer with cross attention. Since this cross attention layer receives the same input at every timestep, the authors introduce a gating mechanism that controls the amount of information to keep at each timestep. Trained on MIMIC-NLE [24], it achieves a reconstruction Bilingual Evaluation Understudy (BLEU)-4 score of 93.9.

4.5.2.2 PEFT

Our AE is trained on text inputs, but it needs to be adapted to incorporate the products of the vision model we want to explain. Recently, the transfer learning paradigm has shifted from fine-tuning entire models to only certain layers. This is especially important in NLE generation, and even more so in post-model methods, where given a pretrained model, we want to obtain explanations for its decision process with the shortest training time possible. Therefore, we experiment with applying several Parameter-Efficient Fine-Tuning (PEFT) techniques to the encoder of our NLE Generator, while keeping its decoder module frozen. These methods are depicted in Figure 4.3 (last three rows). A plethora of PEFT methods has been proposed, not only for single modality use cases [185], [199]–[205], but also for multimodal applications [205]–[207], and in the medical domain as well [208]. Applying PEFT to NLE generation is still an emergent topic, with work only on uni-modal applications [209].

In total, we try five PEFT methods, four of which are built on top of an input configuration similar to that of DPT, i.e., where both image features, prediction vector and associated pathology are given as inputs to the encoder.

LoRA [185] represents, through low-rank decomposition, the model's weight updates into two smaller matrices, which are trained, while the original model weights are kept frozen. After training, both weight matrices are combined.

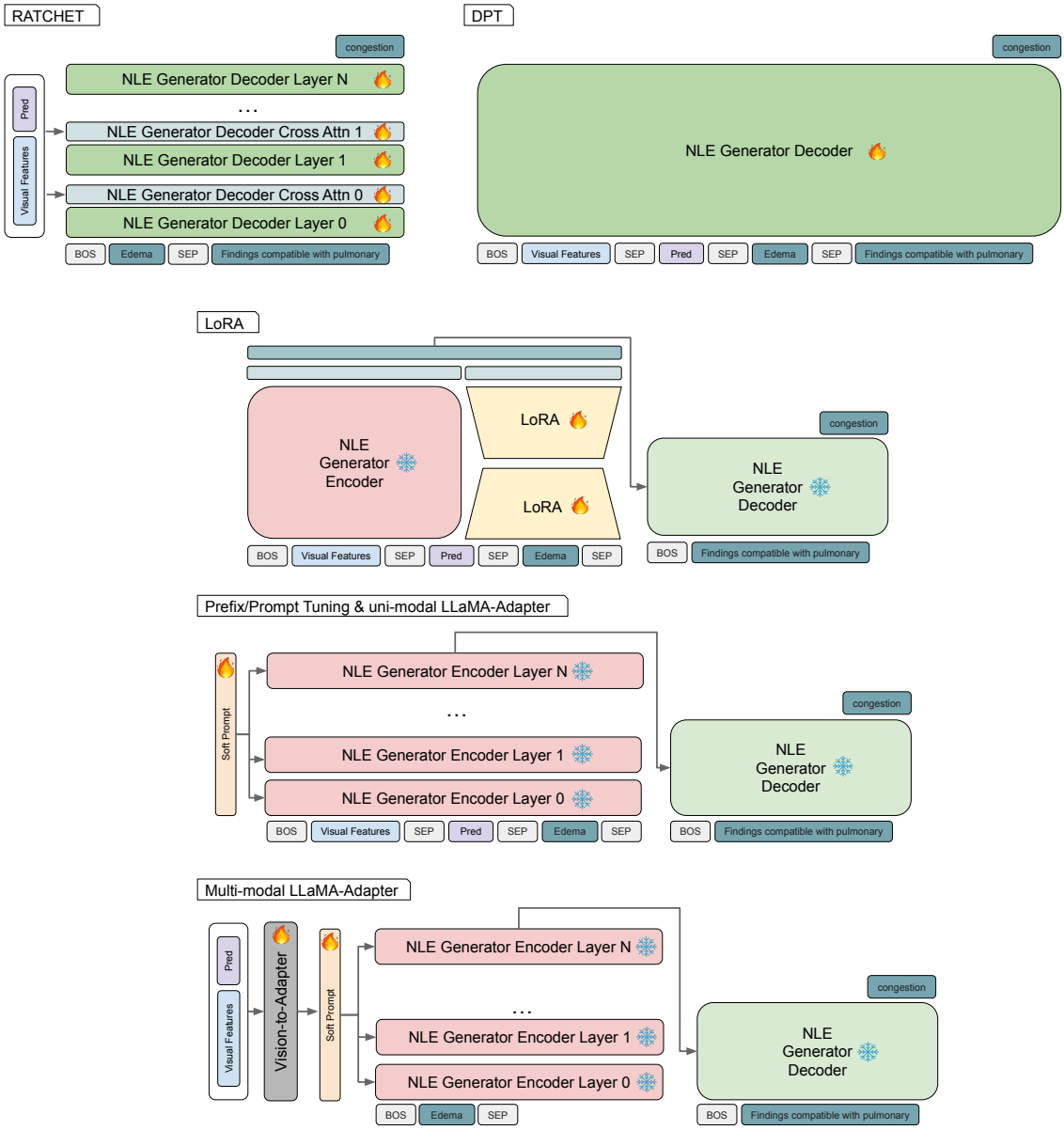


Figure 4.3: The different types of PEFT architectures of the NLE generator. Trainable layers/parameters are represented by the fire icon, while frozen blocks are represented by the snowflake.

Prompt and Prefix Tuning [201], [202] insert a given number of learnable soft prompts into the frozen model. While Prompt Tuning [201] only appends these prompts to the input embeddings, Prefix Tuning [202] applies these prompts to several layers of the frozen model (in our case, to all layers).

LlaMA-Adapter [205] was proposed to efficiently fine-tune a LLM, LlaMA [210], and it differs from Prefix Tuning in the way the learnable adapters are merged into the frozen model. The authors propose to separate the adapter prompts from the word tokens and introduce a zero-initialised gating mechanism that adaptively controls the influence of the adapter prompts during

training. This way, in the early training stages, the influence of the adapter prompt is very small, and it is progressively increased, which guarantees training stability while still ensuring the model is gradually being adapted and simultaneously retaining the knowledge of the original model.

However, none of the aforementioned methods take into account the multimodality of our scenario, as they were developed for uni-modal applications. To tackle this, the authors of LlaMA-Adapter also propose a multimodal version. Instead of providing the image-related information through the input of the model, it is directly added to the adaptation tokens. To ensure the image features have the same dimensions as the adaptation prompts, a Transformer encoder is used, as shown in Figure 4.3 (last row).

4.5.2.3 NLE Generation Loss Function

As previously mentioned, using an AE architecture opens the door to the RL-free adversarial generation of NLEs in low GT availability regimes. To test the viability of such an approach despite having, as we do in this work, GT NLEs for the whole dataset, we experiment with supervising the NLE generation directly in the latent space of the encoder, as this is where the adversarial learning would happen. Thus, in this experiment, we replace the cross entropy loss computed between the discrete output of the decoder and the GT NLE, by the Mean Squared Error (MSE) computed between the encoder latent representation of the GT NLE and the continuous output of the encoder.

4.5.3 Experimental Setup

4.5.3.1 Evaluation Metrics

We follow the evaluation protocol proposed in the MIMIC-NLE paper [24]; several aspects are evaluated, namely the CXR multilabel classification performance (through balanced AUC), and the NLEs' quality (only for correctly predicted instances). Given that there is evidence that automated Natural Language Generation (NLG) metrics correlate badly with the way humans evaluate how similar two NLEs are [175], because two NLEs can be simultaneously correct while having very different syntactic forms, the authors propose the Clinical Evidence (CLEV) score. This score uses the CheXbert labeler [196] to verify if an NLE refers to the same clinical evidence as the GT NLE. Thus, it corresponds to the accuracy regarding clinical evidence. Several NLG metrics are considered, with a bigger importance given to BERTScore [211] and Metric for Evaluation of Translation with Explicit word Ordering (METEOR) [212] since it has been shown that these exhibit the highest correlation with human evaluation [175].

4.5.3.2 Implementation Details

Regarding the pre-training of the AE, the decoder is a single transformer layer, as proposed in the original AUTOBOT paper [198], and the encoder is the CXR-BERT model [37]. CXR-BERT is

a BERT masked language model trained on PubMed abstracts [213], clinical notes from MIMIC-III [214] and MIMIC-CXR [13]. The model is trained with the AdamW optimizer for 1 million steps with a batch size of 64 and a linearly decayed learning rate of 0.001 with 1000 warmup steps. The best model is chosen based on the reconstruction BLEU-4.

For the explanation generation model, we follow the training details described in the MIMIC-NLE paper [24]: for RATCHET, the learning rate is 5×10^{-5}, and 5×10^{-4} for all remaining experiments. All models are trained with the AdamW optimizer for 50 epochs with a batch size of 16 and a linearly decayed learning rate with 1000 warmup steps. We use Early Stopping with a patience of 25.

LoRA parameters follow the original LoRA paper [185], i.e., an alpha of 32 and a dropout of 0.01. For the remaining PEFT methods, the learnable prompts have 10 tokens (as in the original LlaMA-Adapter [205]), and for both LlaMA-Adapter models, the prompts are injected at every layer of the NLE generator encoder. The transformer used in the multimodal LlaMA-Adapter has the following configuration: 2 transformer layers with 4 attention heads, an embedding dimension of 768, and the linear layers have 3072 neurons.

The best model and corresponding hyperparameters are chosen with the same criteria as in the MIMIC-NLE paper [24]: the product of the classification score (balanced AUC), CLEV score, and the average of BERTScore [211] and METEOR [212], computed on the validation set.

We learn the weighting scheme of the classification and text losses, following the methodology proposed in [215]. Thus, our loss function is given by $\mathscr{L} = \mathscr{L}_{clf} + \frac{1}{\sigma^2}\mathscr{L}_{text} + log(\sigma)$, where σ is learned during training and has a starting value of 1.0.

4.5.4 Results and Discussion

Table 4.1 contains the results of the different NLE generation frameworks evaluated in terms of balanced AUC, CLEV and NLG metrics. We also present the number of correctly classified instances (for which the CLEV and NLG metrics are computed), the number of additional trainable parameters introduced, and the overall performance.

Besides the base DenseNet-121 model, we present the results of two of the models of [24], RATCHET and DPT, trained with our loss weighting scheme. The obtained results follow the trends verified in the original paper, i.e., RATCHET improves AUC over the DenseNet-121, while DPT does not, and the former is better overall than the latter.

Regarding our proposals, replacing the decoder-only NLE generator (i.e., DPT) with the AE (Base AE) improves the overall performance: while the AUC slightly decreases, the CLEV and BERT scores increase. However, this comes at a cost: 145M trainable parameters are added, 20M more than DPT introduces.

Uni-modal PEFT techniques slightly decrease or perform on par with the Base AE model in terms of overall performance. However, these methods introduce much less trainable parameters (11M in the worst case). Given the small loss in performance but the high gain in training

Table 4.1: Results obtained on the MIMIC-NLE [24] test set. Evaluation metrics include the balanced AUC, the CLEV score, and several NLG metrics. CLEV and NLG metrics are only computed for correctly classified instances. This value is presented in column # instances, for which the total is 835 (one NLE can explain more than one prediction). The # params column corresponds to the number of trainable parameters (in millions, M) relative to the 7M DenseNet-121 model. The last column shows the overall performance, given by the product of the AUC, the CLEV score, and the average of the BERT and METEOR scores. In bold are the best results for each metric across all models, while underlined are the best results among our models.

Model	AUC	CLEV	BERTS	METEOR	BLEU-1	BLEU-4	ROUGE	CIDEr	SPICE	# instances	# params. (↓)	overall
DenseNet-121	64.45										7M	
RATCHET	**65.19**	74.81	77.49	13.28	19.92	3.933	21.53	37.71	20.04	540	+90.5M	**221337**
DPT	63.24	73.90	77.97	13.57	21.02	**4.364**	**22.65**	**40.67**	19.25	498	+125M	213903
Base AE	63.01	74.19	**78.42**	13.10	15.31	2.497	16.58	11.81	15.02	527	+145M	213915
LoRA AE	63.98	73.99	77.51	**15.32**	20.38	4.002	20.60	24.44	**20.11**	546	+1.1M	219723
Prompt AE	61.31	74.49	77.16	14.43	19.49	3.773	19.76	23.48	18.10	541	**+0.8M**	209145
Prefix AE	**65.19**	67.96	77.56	14.70	20.06	3.739	19.77	21.53	18.84	**568**	+1.2M	204370
LLaMA-Adapt AE	63.87	71.05	77.17	14.61	**23.06**	4.356	21.47	29.74	19.00	532	+1.0M	208247
+ multimodal	64.86	74.63	76.95	14.12	18.77	3.098	18.65	18.40	16.29	536	+11M	220412
+ MSE loss	62.32	**77.78**	78.05	10.04	16.46	2.075	14.20	13.97	9.710	495	+11M	213497

efficiency, PEFT methods are, in fact, suitable for NLE generation for CXR diagnosis, as had already been shown in other domains. Our results also validate the findings of the original LlaMA-Adapter [205], i.e., that the zero-initialised gating mechanism surpasses traditional Prefix Tuning.

The multimodal version of LLaMA-Adapter achieves the best overall results among our proposals, even surpassing the fully fine-tuned model. This confirms our hypothesis that incorporating the visual information throughout the NLE generator layers instead of just at the input is beneficial. Indeed, this is probably the same reason why RATCHET outperforms DPT. This result demonstrates that explainability and task performance are not always at odds with each other and, in fact, explainability might improve classification accuracy.

Finally, we train the best PEFT configuration (multimodal LLaMA-Adapter) directly with MSE on the latent space of the encoder. While the decrease in NLG metrics is expected since supervision now occurs at the sentence level instead of at the token level, the CLEV score significantly increases, which shows that, although the generated NLEs might by syntactically different from the GT NLEs, they contain the correct clinical evidence. Thus, this confirms that our proposed approach is viable, laying the foundation for the RL-free adversarial generation of NLEs.

In Figure 4.4 we present qualitative examples of the NLEs generated by all models. These examples show that, although using only a small percentage of trainable parameters, PEFT approaches produce NLEs that are consistent with NLEs produced by fully fine-tuned models (DPT, RATCHET, and Base AE). Furthermore, they corroborate the known fact that NLG metrics are not good measures of the correctness of a given text: the Base AE model produces the same sentence for all instances, which can be explained by the high imbalance of MIMIC-NLE (evidence labels are predominantly 'Lung Opacity') [24]. For example, consider the last example where 'Atelectasis' is not present in the labels, but the NLE generated by the Base AE mentions that 'There is a new opacity in the right lung base which may represent atelectasis'. However, the majority of the NLG metrics (BERTScore, METEOR, BLEU-4, Recall-Oriented Understudy for



Labels: **Atelectasis** (U), Lung Opacity (P) and Pneumonia (U)

GT: There is new opacity at the right cardiophrenic angle, which may be atelectasis but could also represent pneumonia in the appropriate clinical setting.

RATCHET: Subtle opacity in the left lower lobe is concerning for pneumonia.

DPT: New right lower lobe opacity concerning for pneumonia.

Base AE: There is a new opacity in the right lung base which may represent atelectasis, but cannot exclude pneumonia or aspiration in the right clinical setting.

LoRA: New patchy opacities in the right upper lobe are concerning for pneumonia.

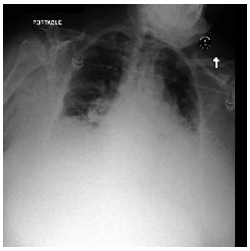
Prompt: Patchy opacity in the left lower lobe is likely due to atelectasis, but infection cannot be excluded in the appropriate clinical setting.

Prefix: There is a new focal opacity in the right upper lobe, concerning for pneumonia or aspiration in the appropriate clinical setting.

LLaMA Uni: Subtle opacity in the left lower lobe is likely atelectasis versus scarring in the appropriate clinical setting, a possible right middle lobe pneumonia cannot be excluded.

LLaMA Mult: Subtle opacity at the right lung base likely represents atelectasis or pneumonia.

LLaMA Mult MSE: There is patchy left basilar opacity likely reflecting atelectasis, but appear free air.



Labels: Atelectasis (U), **Edema** (U), Lung Opacity (P), Pleural Effusion (U) and Pneumonia (U)

GT: Retrocardiac opacity which may be due in part to pulmonary edema and atelectasis limits evaluation of the left lower lung, and superimposed pneumonia cannot be excluded.

RATCHET: There is prominence of the pulmonary vasculature consistent with mild pulmonary edema.

DPT: There is likely due to pulmonary edema.

Base AE: There is a new opacity in the right lung base which may represent atelectasis, but cannot exclude pneumonia or aspiration in the right clinical setting.

LoRA: There is mild interstitial prominence, which could reflect mild pulmonary edema, but not entirely excluded.

Prompt: There is mild interstitial prominence of the hila bilaterally, consistent with pulmonary edema.

Prefix: There is persistent bilateral perihilar opacities, likely representing asymmetric edema.

LLaMA Uni: There is persistent mild interstitial edema.

LLaMA Mult: There is increased interstitial markings in the left lower lobe, which may represent a component of pulmonary edema and atelectasis, less likely pneumonia.

LLaMA Mult MSE: There is hazy associated opacity in the left base likely representing atelectasis.



Labels: Lung Opacity (P) and **Pneumonia** (P)

GT: A new right lower lobe airspace opacity is likely due to aspiration pneumonia.

RATCHET: Right lower lobe consolidation is worrisome for pneumonia.

DPT: New right upper lobe opacity concerning for pneumonia.

Base AE: There is a new opacity in the right lung base which may represent atelectasis, but cannot exclude pneumonia or aspiration in the right clinical setting.

LoRA: There is persistent consolidation in the right upper lobe, which may represent aspiration or pneumonia in the appropriate clinical setting, worsening of the right lung base, and could be due to aspiration or pneumonia

Prompt: New opacification at the right lung base is compatible with pneumonia.

Prefix: There is a new focal opacity in the right upper lobe, concerning for pneumonia or aspiration.

LLaMA Uni: The opacity in the right upper lobe is concerning for pneumonia or aspiration, and less likely atelectasis.

LLaMA Mult: There is a new heterogeneous opacification in the right mid and lower lung zones concerning for pneumonia.

LLaMA Mult MSE: Given that the right basilar opacity is concerning for pneumonia, less likely atelectasis given the clinical setting.

Figure 4.4: Illustrative examples of NLEs generated by all tested models for correctly predicted images. For each image we provide the GT labels, where: U - uncertain and P - positive. The labels highlighted in bold correspond to the disease being explained in the NLE.

Gisting Evaluation (ROUGE), and Semantic Propositional Image Caption Evaluation (SPICE)) for the Base AE surpass those of the MSE loss trained model. On the other hand, both versions of the multimodal LlaMA-Adapter present higher CLEV scores, indicating that they contain the correct clinical evidence, even though the syntax of the generated NLEs might deviate more from the GT NLEs.

4.5.5 Summary and Conclusions

We build upon the seminal work of Kayser *et al.* [24], which tackled the problem of generating NLEs for CXR diagnosis, by replacing the decoder-only NLE generator by an AE, coupled with PEFT techniques to drastically reduce the number of trainable parameters while maintaining competitive performance. We show that using the multimodal-aware LLaMA-Adapter [205] achieves the best overall results, even outperforming the fully fine-tuned AE model. We also empirically demonstrate that directly supervising the NLE generation process on the encoder latent space at the sentence level is viable, thus consolidating this approach as a first step towards the RL-free adversarial generation of NLEs when no (or few) GT NLEs are available.

4.6 Weakly Supervised Approach

Having shown that directly supervising the NLE generation process on the encoder latent space at the sentence level is viable, we move on to tackle the end-goal of this chapter: the weakly supervised generation of NLEs, with the introduction of *WeNLEX*.

4.6.1 Proposed Model and Loss Functions

Figure 4.5 presents the overall architecture of *WeNLEX* and its loss functions that promote the following necessary properties of an NLE [25], [162], [164], [216]:

- plausibility, i.e., sound coherent and logical to a human being
- faithfulness, i.e., reflect the model's decision process
- image-relevance, i.e., be specific to a given image

4.6.1.1 Plausibility

In fully supervised NLE generation architectures, plausibility is achieved by approximating generated NLEs with GT NLEs via, e.g., the cross entropy loss. In our case, to be able to do this in a weakly supervised manner (i.e., without a one-to-one correspondence between generated and GT NLEs), we build a database with a fixed number of GT NLEs per diagnosis label. In practice, this database does not have the GT NLEs in textual form, but their embeddings given by the pretrained and now frozen NLE Generator Encoder. Therefore, this becomes a distribution matching problem

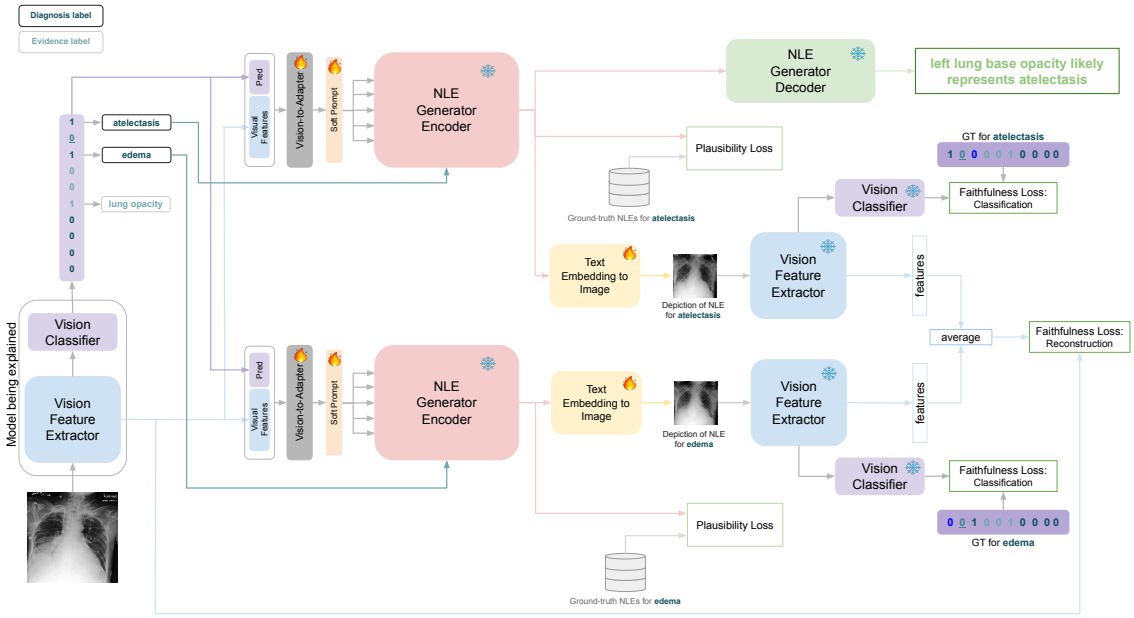


Figure 4.5: Architecture of *WeNLEX*, a weakly supervised model that generates NLEs for a multi-label X-ray classifier. For each predicted diagnosis (e.g., atelectasis, edema), it produces an NLE (only the atelectasis NLE is shown). A pretrained, frozen text-only Encoder–Decoder is adapted with soft prompt tuning (*Vision-to-Adapter* and *Soft Prompt*) to take as input the image features, the entire prediction vector (including diagnosis and evidence labels), and the textual label of the diagnosis being explained. The *NLE Generator Encoder* outputs an NLE embedding, which is compared to GT NLE embeddings for that diagnosis to ensure plausibility (*Plausibility Loss*). Each NLE is also given to a *Text Embedding to Image* model, which generates an image depicting its content. This image is then processed by the MBE to extract features. To enforce faithfulness, the average of these features across all NLEs for an image is compared with the original image features (*Faithfulness Loss: Reconstruction*). Finally, each NLE must recover the MBE's original diagnosis prediction: the MBE's output for the generated image/NLE is compared against the original prediction (*Faithfulness Loss: Classification*). Trainable layers/parameters are represented by the fire icon, while frozen blocks are represented by the snowflake.

between the generated and GT NLEs. We experiment with two ways to tackle this: adversarial learning and Maximum Mean Discrepancy (MMD) minimization.

In the first approach, we use a Wasserstein GAN with Gradient Penalty (WGAN-GP) [217], [218]. WGAN-GP has been proposed to mitigate the mode collapse problems of traditional Generative Adversarial Networks (GANs) and is composed of:

1. the generator, g_{θ_2}, generates an NLE for a predicted diagnosis conditioned on the visual features and the prediction of the vision classifier being explained (f_{θ_1}). It learns to minimize the Wasserstein distance between the generated and real NLE embedding distributions by producing NLEs that the critic will score highly as belonging to the real data distribution.

2. the discriminator (or critic), d_{θ_3}, receives the generated NLE embedding and learns to estimate the Wasserstein distance between the real and generated NLE embedding distributions.

Just like in the traditional GAN paradigm, the generator and discriminator play a min-max game, given by:

$$\min_{g} \max_{d \in \mathscr{D}_{\text{1-Lip}}} \mathbb{E}_{e \sim \mathbb{P}_r}[d(e)] - \mathbb{E}_{\hat{e} \sim \mathbb{P}_g}[d(\hat{e})] \tag{4.1}$$

where g is the NLE generator, e is the GT NLE embedding, $\hat{e}$ is the generated NLE embedding, d is the discriminator, $\mathscr{D}_{\text{1-Lip}}$ is the set of all discriminators that are 1-Lipschitz continuous, P_r is the distribution of real NLE embeddings, and P_g is the distribution of generated NLE embeddings.

The discriminator loss is given by:

$$\mathscr{L}_d = \mathbb{E}_{\hat{e} \sim \mathbb{P}_g}[dD(\hat{e})] - \mathbb{E}_{e \sim \mathbb{P}_r}[d(e)] + \lambda \, \mathbb{E}_{\tilde{e} \sim \mathbb{P}_{\tilde{e}}}\left[\left(\|\nabla_{\tilde{e}} d(\tilde{e})\|_2 - 1 \right)^2 \right] \tag{4.2}$$

$$\tilde{e} = \alpha e + (1-\alpha)\hat{e}, \quad \alpha \sim \mathscr{U}[0,1],\ e \sim \mathbb{P}_r,\ \hat{e} \sim \mathbb{P}_g \tag{4.3}$$

where $\tilde{e}$ is the interpolation between the GT (e) and the generated ($\hat{e}$) NLE embeddings, and α is an interpolation factor sampled from a uniform distribution $\mathscr{U}$ on the interval $[0,1]$.

The generator loss is given by:

$$\mathscr{L}_g = -\mathbb{E}_{\hat{e} \sim \mathbb{P}_g} d(\hat{e})] \tag{4.4}$$

Although WGAN-GP works well in theory, in practice it might be difficult to make it stabilize. Moreover, it introduces additional training parameters due to the discriminator. Therefore, we also experiment with MMD minimization.

MMD measures the distance between two distributions through the distance between their embeddings in the Reproducing Kernel Hilbert Space (RKHS):

$$\begin{aligned} \widehat{\text{MMD}}^2(X,Y) &= \mathbb{E}_{i,j}[k(x_i,x_j)] + \mathbb{E}_{i,j}[k(y_i,y_j)] - 2\,\mathbb{E}_{i,j}[k(x_i,y_j)] \\ &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j) \end{aligned} \tag{4.5}$$

where X is the generated NLE embeddings distribution, Y is the GT NLE embeddings distribution, x_i and x_j (y_i and y_j) are samples from distribution X (Y), m (n) is the number of samples in X (Y), and k is a Gaussian kernel defined by $k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$. The kernel computes all pairwise Euclidean distances between $X \cup Y$.

In practice, we compute MMD per-label and then average over all labels in the batch, so for each MMD computation X are the NLEs generated for a given label, while Y are all of the GT NLEs for that label present in the NLE database.

4.6.1.2 Faithfulness

In addition to being plausible, NLEs should provide enough detail to allow the reconstruction of their corresponding images. Following [219], we do not attempt to reconstruct images in pixel space but instead in the feature space of the MBE. To achieve this, each generated NLE is passed to a module that converts its text embeddings into images. The latter are given to the MBE and the L2 distance between the original and the reconstructed image features is computed. Given the multilabel nature of our primary classification task, there can be several NLEs per image, each referring to a specific pathology. However, the features of the original image contain information on all predicted pathologies (i.e., the features are not disentangled by pathology). So, we actually measure the distance between original image features and the average of all features obtained from each generated image, i.e., from each generated NLE for a given input image (c.f. Fig 4.5).

Additionally, for an NLE to be faithful it also needs to be able to produce the same decision that gave rise to it in the first place. Although this is somewhat implicitly covered with the feature reconstruction loss detailed above, to ensure further disentanglement of the NLEs of a given input image, we also compute the classification loss for each generated image. However, two things are important to mention regarding the target of this classification loss: i) it is not the GT of the original image, but the prediction vector of the classifier being explained (as previously explained the NLEs need to be faithful to the model, not to the human annotations), and ii) it is not the full prediction vector of the classifier, but only the prediction vector for that specific diagnosis (the evidence labels are the same for all targets, since there is no way of distinguishing which evidence label led to each diagnosis prediction).

Finally, in the in-model scenario, we ensure that the original classifier is not updated directly when it is used to compute the feature reconstruction and classification losses, i.e., it is updated but only through the backpropagation path through the Text Embedding to Image model and the NLE Generator Encoder. To achieve this, we maintain a frozen copy of the classifier to be used just for the feature extraction needed for the reconstruction and classification losses. Since the original classifier is being updated at every iteration and that could cause training instability, we only update the copy with the newest parameters every 1000 steps.

4.6.1.3 Image-relevance

Image-relevance is guaranteed in two ways: by giving the visual features extracted by the MBE to the NLE Generator Encoder, and through the feature reconstruction loss, as explained above.

4.6.1.4 Overall Loss Function

WeNLEX is trained with a combination of the three aforementioned loss functions: i) $\mathscr{L}_{nle\_plaus}$, the plausibility loss (either adversarial or MMD), ii) $\mathscr{L}_{nle\_recons}$, the image feature reconstruction loss, and iii) $\mathscr{L}_{nle\_clf}$, the NLE classification loss. In the in-model case, it also includes $\mathscr{L}_{img\_clf}$, the multiclass multilabel image classification loss, since the classifier is being trained alongside the NLE generation (i.e., in this case *WeNLEX* can be considered a self-explanatory model).

The overall loss function uses the automatic loss weighting method of Cipolla *et al.* [215], thus resulting in the following equations:

$$\mathscr{L} = \frac{1}{2\sigma_1^2}\mathscr{L}_{plaus} + \frac{1}{2\sigma_2^2}\mathscr{L}_{nle\_clf} + \frac{1}{2\sigma_3^2}\mathscr{L}_{nle\_recons} + +log(1+\sigma_1^2) + log(1+\sigma_2^2) + log(1+\sigma_3^2) \tag{4.6}$$

For the in-model case, the overall loss function includes the term $\frac{1}{2\sigma_4^2}\mathscr{L}_{img\_clf} + log(1+\sigma_4^2)$. The σ_i are learnable parameters with an initial value of 1.

4.6.2 Experimental Setup

4.6.2.1 Implementation Details

Unless otherwise stated, all models are trained with the AdamW optimizer for 50 epochs with a batch size of 16 and a linearly decayed learning rate of 5×10^{-4} with 1000 warmup steps. The final model is chosen based on the lowest validation loss.

Multilabel Classifier Following previous work on the MIMIC-NLE dataset [24], [26], the vision classifier being explained is the DenseNet-121 [88]. In the *post-hoc* experiments, the classifier is pretrained on the images of the MIMIC-NLE dataset with a class-weighted cross entropy loss, obtaining an AUC of 65.13%.

Text Autoencoder The text AE (NLE Generator Encoder and NLE Generator Decoder blocks of Figure 4.5) is the same as in Rio-Torto *et al.* [26], with the only difference being related to the pretraining dataset. In [26] the encoder was the CXR-BERT model [37], which is a masked language BERT-based transformer model trained on PubMed abstracts [213], clinical notes from MIMIC-III [214] and MIMIC-CXR [13]. Since in this work we are operating under the weakly supervised setting, we assume we do not have access to all MIMIC-NLE sentences. Considering that MIMIC-NLE has been derived from MIMIC-CXR, we wanted to make sure our encoder never had access to MIMIC-CXR. Thus, we take the PubMedBERT [220][1] (BERT model trained from scratch on PubMed abstracts only) and fine-tune it for 500k steps and an initial learning rate of 2×10^{-5} on sentences of the Interpret-CXR dataset [221] (from which we exclude MIMIC-CXR). Afterwards, we follow the procedure described in [26], [198] and train, still on Interpret-CXR, the single-layer transformer decoder with a denoising AE objective for 1M steps, batch size of 64, and an initial learning rate of 1×10^{-3}. It achieves a reconstruction BLEU-4 score of 85.9.

Discriminator When using adversarial learning as the plausibility loss the whole architecture of *WeNLEX* except the discriminator is considered the generator (Vision-to-Adapter, NLE Generator Encoder, Text Embedding to Image model, etc). The discriminator is implemented as a lightweight feed-forward network that receives the NLE and a diagnosis-specific embedding (the

[1] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

latter is obtained by passing each diagnosis label through the frozen NLE Generator Encoder). The two 768-dimensional vectors are concatenated and passed through a series of fully connected layers with progressively reduced dimensionality (2×768, 512, 256, and 1). The hidden layers are followed by LeakyReLU activations ($\alpha = 0.2$). As is commonly done when training Wasserstein GANs (WGANs), for every 5 discriminator updates the generator is updated once.

Text Embedding to Image Model The Text Embedding to Image model converts the 768-dimensional NLE embeddings produced by the NLE Generator Encoder to 224×224 grayscale images. The embedding is first projected into a $512 \times 14 \times 14$ feature map using a fully connected layer with ReLU activation. It is then upsampled through four transposed convolutional layers, each doubling spatial resolution while reducing channels ($512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 1$). Batch normalization and ReLU are applied after each intermediate layer, and a final Tanh activation normalizes the output image to $[-1, 1]$.

NLE Database To build the GT NLE database we randomly sample n NLEs per diagnosis label, excluding NLEs that explain more than one diagnosis (e.g., "New opacification at the right lung base could be either atelectasis or developing pneumonia."), in order to obtain NLEs as specific to each diagnosis as possible.

4.6.2.2 Evaluation Metrics

Evaluating text generation tasks, e.g., image captioning, is notoriously difficult given the inherent variability and ambivalence of natural language, and still somewhat of an unsolved problem in the field [168], [175]. Evaluating NLEs is even harder because, e.g., compared to image captioning, another variable needs to be taken into account: the MBE.

Plausibility Most works on NLE generation focus only on measuring the plausibility of NLEs by comparing them to human annotated NLEs via common NLG metrics, such as BLEU [222], METEOR [212], BERTScore [211], etc. Since it has been shown that BERTScore correlates higher with human evaluation [175] and since our NLEs are from the medical domain, we compute their plausibility via BERTScore with the CheXbert [196] model. This score, the CheXbert Score (CXBS), is computed only for correctly predicted NLEs (correct diagnosis and evidence labels).

Simulatability Another way of evaluating NLEs (in fact, any other kind of explanation) is through their utility to an end-user, which is defined by the concept of simulatability: how well an observer can use explanations to obtain the MBE's prediction [223], [224]. This concept appears in the literature under different names and evaluation procedures [176], [178], [225], e.g., in NLX-GPT [178] the authors call it "Explain-Predict" and give the input question and the explanation to a language model and ask it to answer the question (in the context of a visual question-answering task). Wiegreffe *et al.* [225], in the context of a task with text-only input, measure the additional ability to predict a label that an explanation provides over the input, i.e., the difference between

task performance when an explanation is given together with the input vs. when it is not. We adapt both aforementioned approaches to our case and present the $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$ metrics. They measure if the NLE alone or together with the input image allow a human to reach the same prediction as the MBE. Following previous work [224], we use a model as a proxy for the human observer; we ask the vision-language foundation model CheXagent [115] to answer if a given label is present in the NLE or in both NLE and image. We also compute $\hat{y}|img$ to establish a baseline on the performance of CheXagent on the images alone.

Faithfulness Although several metrics have been proposed to assess the faithfulness of NLEs, they are mostly tailored for settings where the input is text [162], [226], [227]. Both in the realm of visual explanations and NLEs for tasks where the input is text, a popular way to measure faithfulness is by removing parts of the input that the explanation considers important and measure the decrease in the performance of the MBE [103], [104], [227]. Wojciechowski *et al.* [180] adapt this paradigm for the vision-language domain by asking humans to cover parts of the input image containing the decision rationale indicated in the NLE. We do the same but by leveraging the phrase grounding capabilities of CheXagent: we ask it to ground the NLE in the image and then mask the image in the location(s) given by the bounding boxe(s) produced by CheXagent (see Figure 4.6). Following Wojciechowski *et al.* [180] we measure the number of times the multilabel classifier flips its decision when given the masked image and the absolute difference in the logits of the diagnosis originally predicted (we consider both *Uncertain* and *Positive* logits together).

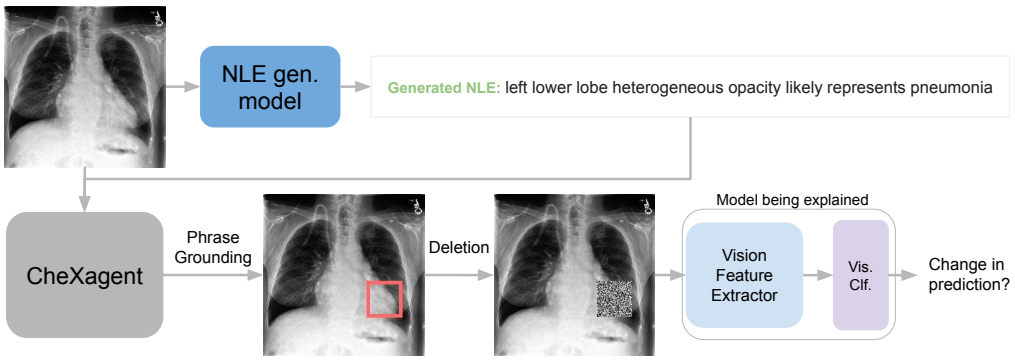


Figure 4.6: Depiction of the deletion faithfulness metric: an image and a generated NLE are given to CheXagent, which grounds the text in the image. The identified regions are occluded, and the masked image is given to the MBE. If the NLE is faithful, occluding these regions should significantly alter the MBE's prediction.

Following the evaluation protocol of the work that originally proposed the MIMIC-NLE dataset [24], we also compute the CLEV score, which uses the CheXbert [196] model to check if two NLEs refer to the same clinical evidence. However, we introduce two modifications: i) given the big imbalance in the evidence labels of the dataset (see Subsection 4.3), instead of reporting the accuracy, we report the macro-averaged F1-score, and ii) we do not compare the evidence of the generated NLE with the evidence of the GT NLE, because the NLE needs to be faithful to the

evidence the multilabel classifier predicted, and this might differ from the GT evidence. An example illustrating this can be found in Figure 4.7, where the GT refers to the *Lung Opacity* evidence label, but the classifier did not predict this, so the generated NLE cannot and should not mention the presence of a lung opacity. Computing the score in this way has another advantage: since we are now only dependent on the input evidence, the CLEV score can be computed for all generated NLEs, and not only for the correctly predicted ones. In fact, apart from the CXBS which directly compares against GT NLEs, all other metrics presented throughout the remainder of this work are computed for all generated NLEs.

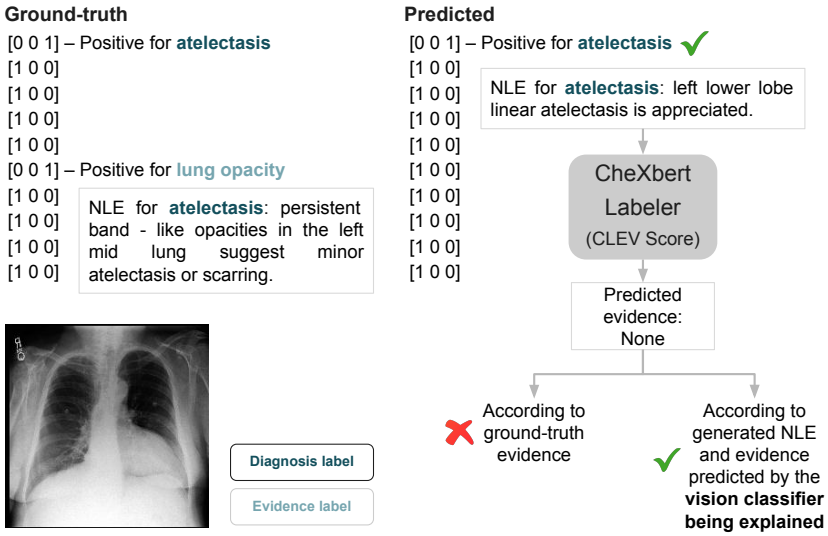


Figure 4.7: Correct CLEV computation: the CheXbert identified evidence is correct if it matches the evidence predicted by the MBE (which is also the evidence used to generate the NLE in the first place), but it is incorrect if judged against the GT evidence. Since we want NLEs faithful to the MBE, the target evidence for computing the CLEV score should be the MBE's predicted evidence.

Diversity Finally, we measure the diversity of the generated NLEs. Although two NLEs for the same disease and similar images should be similar, at the same time they should not be completely equal, since the NLEs also need to be image-relevant. If they were equal one might assume that having generated the same NLEs is due to correlations and biases in the dataset and not due to the reasoning of the MBE [178]. To measure this, we employ the retrieval-based attack proposed in [178]: the 10 most similar images to a query image are retrieved using any feature extraction model (we use MedImageInsight [150]) and *WeNLEX* generates one NLE for each. Then, the cosine distance between these NLEs is computed using Sentence-BERT [228][2]. The lower the distance, the lower the bias, so the better the NLEs will be. To complement this retrieval-based attack we also compute the per-diagnosis Self-BLEU score, i.e., the BLEU-4 score between all NLEs generated for a given diagnosis. The lower the score, the more diverse are the generated NLEs.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

4.6.3 Results

Quantitative results are presented in Table 4.2. We compare *WeNLEX*, in both *post-hoc* and in-model versions (models 5 and 6, respectively), with the fully supervised model of Rio-Torto *et al.* [26] retrained on our modified version of the MIMIC-NLE dataset.

We compare the effect of the proposed loss functions under the *post-hoc* setting, since it is the only setting where we are actually comparing the same NLEs (1038 generated NLEs to be precise, of which 128 have the same diagnosis and evidence labels as their GT counterparts). When comparing the two strategies for promoting plausibility, adversarial learning and MMD minimization (models 2 and 3), it can be seen that MMD minimization increases the CLEV score from 4.876 to 7.601, increases $\hat{y}|NLE$ from 71.77 to 86.89, and increases CXBS from 36.21 to 50.39. Thus, MMD minimization is clearly allowing *WeNLEX* to produce NLEs closer to the GT NLE distribution than the adversarial approach. Moreover, the diversity of the NLEs generated by the adversarial approach is much lower than that of the MMD NLEs (41.58 Self-BLEU vs 10.36), which might indicate that the WGAN-GP might be collapsing to very similar NLEs for each diagnosis. For these reasons, we chose to continue all remaining experiments with the MMD loss.

Introducing the feature reconstruction loss (model 4) increases the CLEV score from 7.601 to 9.303, the percentage of prediction flips from 91.95 to 92.48, and the absolute difference in the prediction from 0.3192 to 0.3236. This increase in faithfulness comes at the cost of slightly less similarity to GT NLEs (CXBS descreased from 50.39 to 47.50), but a significant increase in the diversity of the NLEs (from 10.36 Self-BLEU to 2.671 and from 0.6525 to 0.5258 in the retrieval metric). Therefore, we conclude that, as hypothesised, introducing the feature reconstruction loss improves the faithfulness of the NLEs. However, both $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$ decrease (from 86.89 to 77.84 and from 89.21 to 86.71, respectively). This might simply be due to the fact that the NLEs of model 4 might be more different from the text that the CheXagent model was trained on compared to the NLEs of model 3.

Adding the NLE classification loss (model 5) further increases the CLEV score and substantially increases the $\hat{y}|NLE$ score from 77.84 to 90.46. This comes at a cost of a slight decrease in diversity (e.g., Self-BLEU increases from 2.671 to 4.037), which is to be expected since the classification loss promotes that NLEs for the same disease become similar. However, the increase in CLEV score from 9.303 to 9.401 shows that it brings benefits in terms of ensuring the correct clinical evidence.

In all cases it can be verified that adding the NLE helps CheXagent predict the label better, as its $\hat{y}|img$ accuracy is always lower than $\hat{y}|(img, NLE)$.

In the in-model scenario (model 6) we obtain the best CLEV, $\hat{y}|NLE$, and Self-BLEU metrics, outperforming the fully supervised scenario, using only 5 GT NLEs per diagnosis. It also achieves the best percentage of flipped decisions amongst the weakly supervised model variants.

Both the classifier trained on its own (models 2-5) and the classifier trained together with the explanation generation part of *WeNLEX* improve upon the AUC of the classifier trained together

Table 4.2: Comparison of *WeNLEX* with the fully supervised baseline of [26] in terms of Faithfulness, Simulatability, Diversity, and Plausibility. 🔥 refers to the in-model scenario - the MBE is trained alongside the rest of *WeNLEX*), while ❄ refers to the *post-hoc* scenario (MBE has been pretrained and is frozen during training of *WeNLEX*). "Clf." refers either to the image or the NLE classification loss. "Plaus." refers to the plausibility loss and "Recons." to the feature reconstruction loss. "# NLEs" is the total number of generated NLEs; values in parentheses are for correctly classified cases (diagnosis and evidence). All metrics use the total # NLEs, except CXBS, which uses only the parenthesised values. The ↓ means that for the diversity metrics lower is better. Best results are in bold; second-best are underlined.

Model	Image-Related			NLE-Related											
								Faithfulness			Simulatability		Diversity ↓		Plausibility
									Deletion						
	Clf.	AUC	$\hat{y}\|img$	Plaus.	Recons.	Clf.	# NLEs	CLEV (F1)	Flip (%)	Δp	$\hat{y}\|NLE$	$\hat{y}\|(img,NLE)$	S-BLEU	Retrv.	CXBS
Fully Supervised															
1	🔥	64.08	**83.29**	MSE	✗	✗	1107 (128)	10.37	**97.63**	**0.565**	89.97	88.26	8.785	0.7263	48.00
Weakly Supervised															
2	❄	65.13	82.47	Adv.	✗	✗	1038 (128)	4.876	94.72	0.2671	71.77	89.11	41.58	0.8079	36.21
3				MMD	✗	✗		7.601	91.95	0.3192	86.89	**89.21**	10.36	0.6525	**50.39**
4					✓	✗		9.303	92.48	0.3236	77.84	86.71	**2.671**	**0.5258**	47.50
5					✓	✓		9.401	91.48	0.3328	90.46	88.82	4.037	0.6286	47.35
6	🔥	**67.34**	79.91		✓	✓	1115 (168)	**10.61**	95.65	0.3197	**91.21**	86.91	**1.920**	0.6704	45.38

with the fully supervised pipeline of [26] (65.13/67.34 vs 64.08). When comparing only the *post-hoc* and in-model versions of *WeNLEX* (models 5 and 6) we can see that AUC improves when training the classifier together with the whole model, which shows that, unlike what has been many times stated in the literature [7], [229], adding interpretability does not have to mean a decrease in task performance.

4.6.4 Ablations

4.6.4.1 Layers for Feature Reconstruction

We perform an ablation study on which classifier layer to use for the feature reconstruction loss. In the original Perceptual Loss paper [230], the authors recommend using an intermediate layer for this purpose, since higher layers do not preserve shape as well as lower layers. In our case there does not seem to clearly exist a layer that is better than the others with regards to the several metrics we evaluate, as can be seen in Table 4.3. Using higher layers such as the output of the classification layer or the output of the Global Average Pooling (GAP) layer, leads to less diverse NLEs (Self-BLEU scores of 5.481 and 11.38, respectively). More intermediate layers such as the output of the 4th and 3rd denseblocks lead to lower faithfulness scores (CLEV, $\hat{y}|NLE$, $\hat{y}|(img,NLE)$, and Flip. Lower layers like the 2nd and 1st denseblocks improve the CLEV score, while keeping the diversity higher than with the highest level layers (classifier and GAP). They also provide the highest plausibility (CXBS). Thus, we choose to use denseblock2 for all other experiments in this work, since it achieves similar metrics to denseblock1, but with a decreased computational cost (exactly half the number of features).

Table 4.3: Ablation on the layer used for the feature reconstruction loss. "# Feats" is the dimension of the flattened feature map. "classifier" is the output layer, "gap" the GAP layer, and "denseblock" the outputs of DenseNet layers (with "denseblock1" closest to the input). The underlined layer is used in all other experiments. Best results are in bold; second-best are underlined.

Layer	# Feats.	Faithfulness			Simulatability		Diversity ↓		Plausibility
			Deletion						
		CLEV (F1)	Flip (%)	Δp	$\hat{y}\|NLE$	$\hat{y}\|(img,NLE)$	S-BLEU	Retrv.	CXBS
classifier	30	**9.623**	91.81	0.3377	<u>91.81</u>	**89.31**	5.481	0.6580	46.96
gap	1024	7.993	**92.39**	0.3279	**92.49**	88.34	11.38	0.6927	46.98
denseblock4	50176	8.887	91.62	**0.3483**	86.13	88.05	<u>2.424</u>	<u>0.5907</u>	46.80
denseblock3	200704	8.631	91.16	0.3344	86.03	87.57	**1.992**	**0.5767**	46.46
<u>denseblock2</u>	401408	9.401	91.48	0.3328	90.46	<u>88.82</u>	4.037	0.6286	<u>47.35</u>
denseblock1	802816	<u>9.463</u>	<u>92.09</u>	<u>0.3385</u>	89.31	88.63	3.654	0.6231	**47.41**

4.6.4.2 Size of Ground-truth NLE Database

We also ablate the number of clinician-annotated (i.e., GT) NLEs per diagnosis label present in the NLE database. We test *WeNLEX* with 2, 5, 10, and 20 NLEs per diagnosis label. Each experiment is run with 3 different random seeds and the results (mean and standard deviations) are reported in Table 4.4.

As expected, using only 2 NLEs per diagnosis label achieves the lowest diversity (Self-BLEU of 8.544 and retrieval score of 0.6852). It also achieves the lowest $\hat{y}|NLE$, $\hat{y}|(img,NLE)$, and Δp. Using 5, 10 or 20 NLEs does not yield significant differences, which attests to *WeNLEX*'s stability and robustness without overreliance on the GT NLEs. Given that using 5 NLEs achieves the highest CLEV score and a lower annotation burden than using more NLEs, all other experiments in this work are performed with 5 NLEs per diagnosis label in the database.

Table 4.4: Ablation on the number of NLEs per diagnosis present in the GT NLE database. The underlined # NLEs is used in all other experiments. Best results are in bold; second-best are underlined.

# NLEs	Faithfulness			Simulatability		Diversity ↓		Plausibility
	CLEV (F1)	Deletion		$\hat{y}\|NLE$	$\hat{y}\|(img,NLE)$	S-BLEU	Retrv.	CXBS
		Flip (%)	Δp					
2	8.663 (0.9852)	**92.76** (0.2796)	0.3530 (0.0342)	85.61 (4.097)	85.64 (0.6675)	8.544 (1.727)	0.6852 (0.0382)	**47.40** (1.672)
<u>5</u>	**9.615** (1.157)	92.14 (0.2282)	0.3539 (0.0324)	91.78 (1.404)	**88.63** (0.4199)	3.560 (1.551)	0.6331 (0.0158)	44.88 (3.371)
10	8.396 (0.1789)	92.33 (0.4732)	**0.3749** (0.0016)	<u>92.29</u> (3.424)	<u>88.34</u> (0.3474)	**1.827** (0.0461)	<u>0.6276</u> (0.0348)	45.22 (1.422)
20	<u>8.672</u> (1.735)	<u>92.52</u> (0.4358)	<u>0.3616</u> (0.0106)	**92.77** (2.569)	88.05 (0.4199)	<u>2.090</u> (0.6986)	**0.5976** (0.0400)	<u>46.45</u> (0.9763)

4.6.5 Audience-Adaptable NLEs

By not needing one NLE per image per diagnosis and simply using a few NLEs per diagnosis, not only does *WeNLEX* significantly lower annotation costs while keeping the generated NLEs faithful, but it easily allows for the adaptation of the style of the generated NLEs; one simply has to switch the NLEs in the database. This tackles another very important desirable property of NLEs: adaptability to the target audience [160], [231].

To test *WeNLEX*'s ability to generate NLEs for a different target audience (i.e., not radiologists) we ask GPT4-o [232] to convert our NLE database of 5 NLEs per diagnosis label into sentences understandable by lay people, following the prompt used in Zhao *et al.* [233], and then use this converted database to train another version of *WeNLEX*.

The results are presented in Table 4.5 and Figure 4.8. Naturally the CLEV score decreases, since it uses the CheXbert [196] labeler, which has not seen non-medical sentences during its training. Interestingly, although $\hat{y}|NLE$, $\hat{y}|(img, NLE)$, Flip and Δp all use CheXagent [115], the last two do not seem to be as affected by the layman NLEs as the first two. This might be due to the fact that the location of the findings is given in more similar terms to the original NLEs than the findings themselves (e.g., in the first example of Figure 4.8 the location cue "at the left base" has been switched to "of the left lung in the lower field of the lung", while the evidence and diagnosis, "basilar opacity" and "atelectasis or scarring", have been switched to "widened areas" and "barely visible"), such that CheXagent is still able to perform the visual grounding necessary for Flip and Δp, but is no longer able to identify the diagnosis label, which is needed for $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$.

As expected, the diversity is similar in both scenarios and the plausibility decreases, since the layman NLEs are significantly different from the clinician-annotated NLEs.

Finally, we evaluate the readability of the NLEs (both generated and the ones in the database). For this we use the `textstat` Python package[3] and the `text_standard` consensus metric, which computes the school grade level required to understand a given text, based on several tests (Flesch Reading Ease formula, Flesch-Kincaid Grade Level, Fog Scale, etc). In agreement with the proposed changes, using the simplified NLEs increases the readability (the original NLEs are readable at college level, 13.21, while the simplified NLEs are readable from 10th grade, 10.42). Thus, *WeNLEX* is, in fact, able to adapt to different target audiences.

Table 4.5: Comparison of *WeNLEX* trained on clinician vs. simplified layperson NLEs. The simplified version shows higher readability (lower required grade level). "Gen." denotes generated NLEs, and "DB" the GT NLE from the medical and non-medical databases.

Database	Faithfulness			Simulatability		Diversity ↓		Plausibility	Readability ↓	
	CLEV (F1)	Deletion		$\hat{y}\|NLE$	$\hat{y}\|(img, NLE)$	S-BLEU	Retrv.	CXBS	Gen.	DB
		Flip (%)	Δp							
Medical	9.401	91.48	0.3328	90.46	88.82	4.037	0.6286	47.35	13.21	15.00
Layman	4.025	91.43	0.2842	64.35	84.30	4.534	0.5295	39.49	10.42	10.46

4.6.6 Summary and Conclusions

We proposed *WeNLEX*, a flexible weakly supervised framework for generating NLEs for multilabel CXR classification. Through image feature reconstruction and distribution matching, its explanations are both faithful to the classifier's reasoning and clinically plausible. *WeNLEX*'s

[3] https://github.com/textstat/textstat



Atelectasis

Ground-truth: there is persistent left basilar opacification, likely reflecting compressive atelectasis.

WeNLEX: linear basilar opacity is likely present with atelectasis or scarring at the left base.

Layman WeNLEX: there may be several widened areas of the left lung in the lower field of the lung being barely visible.

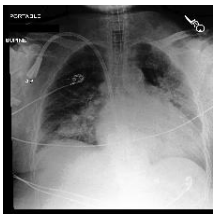


Pleural Effusion

Ground-truth: a hazy opacity over the left lung base suggests a layering pleural effusion.

WeNLEX: left retrocardiac opacity, compatible with effusion / infiltrate.

Layman WeNLEX: there is a marked abnormal area to the chest is in the left lung, which may indicate some increase in the pleural fluid.



Pneumonia

Ground-truth: an underlying infectious infiltrate can't be excluded.

WeNLEX: new right middle lobe opacification, likely present aspiration or pneumonia.

Layman WeNLEX: a hazy area in the right lung is more opacified and could be a very early manifestation of infection.

Figure 4.8: Qualitative examples of *WeNLEX* for three different diagnosis, comparison with GT NLEs, and with the layman version of *WeNLEX*, in which the generated NLE are simplified to adapt to a non-medical audience.

versatility allows it to work in a *post-hoc* or an in-model manner. Beyond explainability, the in-model version of *WeNLEX* increases AUC by 2.21%, demonstrating that explainability and task performance can go hand in hand. *WeNLEX* also allows explanations to be tailored to different audiences, from clinicians to lay users, highlighting its broad applicability in real-world medical settings.

Chapter 5

Vision-Language Explanations

The first research work described in this chapter [29] was conducted in collaboration with Meghal Dani, and Stephan Alaniz, under the supervision of Professor Zeynep Akata, during a six month research visit in the Explainable Machine Learning Group of Professor Zeynep Akata, at the time at the University of Tuebingen (now at Helmholtz Munich). The author of this thesis was mainly involved in the implementation, conceptualization and running of experiments, results analysis and discussion, and preparation of the scientific publication.
The second work [35] was conducted in collaboration with Bruno Coelho, under the supervision of Professors Jaime S. Cardoso, and Luís F. Teixeira. The author of this thesis was involved in all steps of the process. Bruno Coelho offered crucial help in discussing interpretability metrics, the results, and proofreading the manuscript.
The results of these works have been disseminated in two articles in international conferences:

[29] Isabel Rio-Torto*, Meghal Dani*, Stephan Alaniz, and Zeynep Akata, "*DeViL*: Decoding Vision features into Language" Proceedings of the DAGM German Conference on Pattern Recognition (GCPR), pp. 363-377, 2023

* Equal contribution

Code: https://github.com/ExplainableML/DeViL

[35] Isabel Rio-Torto, Bruno Coelho, Jaime S. Cardoso, and Luís F. Teixeira, "Understanding Chest X-ray Vision Representations with *DeViL*" accepted at the IEEE International Symposium on Computer-Based Medical Systems (CBMS)

Having explored visual and textual explanations, we end this thesis by combining both, with the proposal of *DeViL*, a model capable of generating image descriptions and open-vocabulary saliency maps, by training a lightweight translation network on image-caption pairs (i.e., a weak supervision signal w.r.t. the visual explanation generation task). We first describe the original model, which was tested on natural images, and then use it to analyse different CXR encoders.

5.1 Context and Motivation

Throughout this thesis we have shown that there are different ways of explaining the decisions of neural networks. Visual explanations are by far the most popular and NLEs, although less investigated, are also seen in the literature. However, works that simultaneously combine both are rare. Moreover, most XAI methods in the literature, and all methods used and proposed in this thesis so far, require task-specific classifiers (e.g., require a prediction from which to backpropagate from), limiting their applicability, e.g., for analysing vision encoders independently of downstream tasks. Such analysis can be particularly useful when comparing different representation learning strategies. Additionally, these methods can only explain the predefined set of classes for which the MBE was trained.

Recent work has begun to address this latter limitation. For instance, the Narrowing Information Bottleneck (NIB) method [234] is able to generate open-vocabulary saliency maps (i.e., saliency maps can be generated for any word or sentence, not being limited to the set of predefined labels of a downstream task) for Vision-Language Models (VLMs) by measuring changes in image–text mutual information. While this approach removes the need for a classification head, it is specifically designed for CLIP-based models [235], [236] and therefore cannot be applied to vision-only encoders. Furthermore, while it is able to provide visual explanations, it lacks the textual explanation modality.

Thus, in this work, we propose *Decoding Vision features into Language (DeViL)*, a framework that translates visual features of any vision encoder into natural language through image-caption alignment. By employing a general purpose language model, *DeViL* inherits rich generalization properties, allowing it to produce open-vocabulary saliency maps. Additionally, *DeViL* requires no task-specific heads, since it only uses the frozen vision encoder. Therefore, it tackles all aforementioned challenges: i) combines visual and textual explanation modalities; ii) can be applied to any vision encoder, not only VLMs like NIB [234]; iii) does not require task-specific heads; and iv) is not limited to the set of predefined classes of the MBE, i.e., it can generate open-vocabulary saliency maps.

Figure 5.1 illustrates the explanations obtained by *DeViL* when applied to an ImageNet-trained ResNet50 model. If one wants to understand what the network encodes at a specific location of the image, *DeViL* translates the nearest image feature of a layer into language, e.g., at the bottom left the network encodes “pizza on a wooden table” at the last convolutional layer, and “a plate of food at the table” at a lower layer. Subsequently, one can inspect the network in a broader sense, by querying the keywords “pizza” and “food” at the respective layers for a more complete explanation. Evidently, layer 49 understands fine-grained concepts while an earlier layer produces a general word “food”. Similar observations have been previously discovered [89], [237]–[239] and *DeViL* allows us to expose these findings through explorative analysis to users.

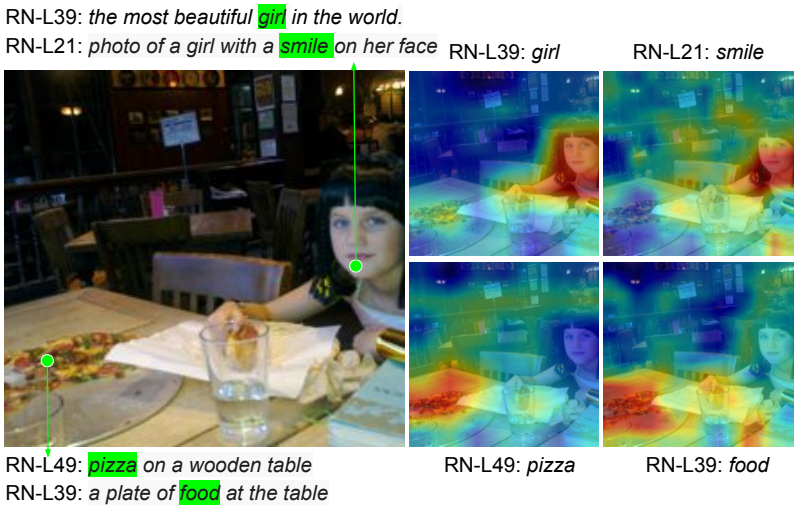


Figure 5.1: Our *DeViL* model generates textual descriptions for feature maps of any arbitrary vision backbone. We show the attribution maps for *girl*, *smile*, *pizza* and *food* concepts generated by the layer 49 (RN-L49), layer 39 (RN-L39) and layer 21 (RN-L21) of a ResNet-50 backbone respectively. We select these words from the sentences that are generated for these layer outputs (highlighted in green).

5.2 Related Work

5.2.1 Inherently Interpretable Models

Apart from *post-hoc* explanation methods, inherently interpretable models or intrinsic explanation methods try to make modifications to the network architecture or loss function to make the deep model more interpretable. For instance, induced alignment pressure on the weights during optimisation either in a linear [3], [240]–[242] or non-linear manner [243], [244] has been shown to produce more interpretable models. While these methods produce more interpretable models they require adaptation and re-training of existing models and cannot always recover the original task performance. With the increasing size of models, this is not always a scalable solution. *Post-hoc* methods trade-off the guaranteed faithfulness of the explanation with broader applicability.

5.2.2 *Post-hoc* Saliency

Saliency attribution is most commonly produced through perturbations [103], [245], gradients [112], activations [238], [246], [247] or the combination of those [5], [248]. CAM [247] and its successors, such as Grad-CAM [5] and Grad-CAM++ [248] are one of the most popular attribution methods that make use of network activations, similar to our approach, but additionally use the gradients of the target class. We set ourselves apart from these methods by also providing directly related textual descriptions for the saliency maps.

5.2.3 Natural Language Explanations

NLEs are another alternative for explaining a network in a human understandable way. One of the first works on NLEs [164] used a Long Short-Term Memory (LSTM) network with a RL-based loss to guide generated image descriptions into more class discriminative explanations. More recently, the availability of LLMs has also seen them applied to NLE generation. These approaches are usually divided into predict-explain and explain-predict paradigms: in the first, a language model is fine-tuned with features and predictions from a pretrained vision model [174], [175], while in the latter a single model is jointly trained to explain and predict [176]–[178].

While the aforementioned approaches only provide a single NLE for the whole image, Bau *et al.* [249] attempts to label individual highest activated neurons in a network by correlating them with pixel-wise annotations. Building upon that, Hernandez *et al.* [33] collect the MILANNOTATIONS datasets where humans labeled sets of images that showed high activations for individual neurons of different network architectures. By training a vision-language model on this data, individual neurons of different layers can be described in natural language.

In contrast, *DeViL* can generate both types of textual descriptions (global and local), without requiring a dataset of human annotated descriptions per neuron or layer. Nonetheless, we show that when our model is trained on MILANNOTATIONS, we obtain a higher correspondence with human descriptions.

5.2.4 Captioning Models

Image captioning models [235], [250]–[253], on the other hand, use LLMs to generate free-form captions for the input image. State-of-the-art in captioning is obtained by large-scale pre-training and subsequently fine-tuning on target datasets [252], [253]. On the other hand, lightweight models such as ClipCap [250] perform competitively by relying on pretrained vision and language models requiring significantly fewer Graphics Processing Unit (GPU) hours to train. *DeViL* is similar to ClipCap in that both models make use of a translation network between image features and the language model prompt. While ClipCap is capable of generating high-quality captions for images, it cannot be directly used to interpret a model's internal representations. In contrast, by extracting features from multiple layers of the vision model and incorporating dropout, *DeViL* generalizes to more fine-grained descriptions and improves all captioning metrics on CC3M [28].

5.3 *DeViL*: Decoding Vision features into Language

5.3.1 Model Architecture

Given a pretrained vision model g, our goal is to decode the features of several of its L layers into natural language describing what these features encode. These descriptions are specific to the network's activations corresponding to an input image. Let $g_l(x)$ denote the feature map of the lth-layer of g for input image x with spatial size $H_l \times W_l$. We propose *DeViL*, a network that is trained to decode an individual feature vector $g_l(x)_{i,j}$ of an arbitrary spatial location i, j into a

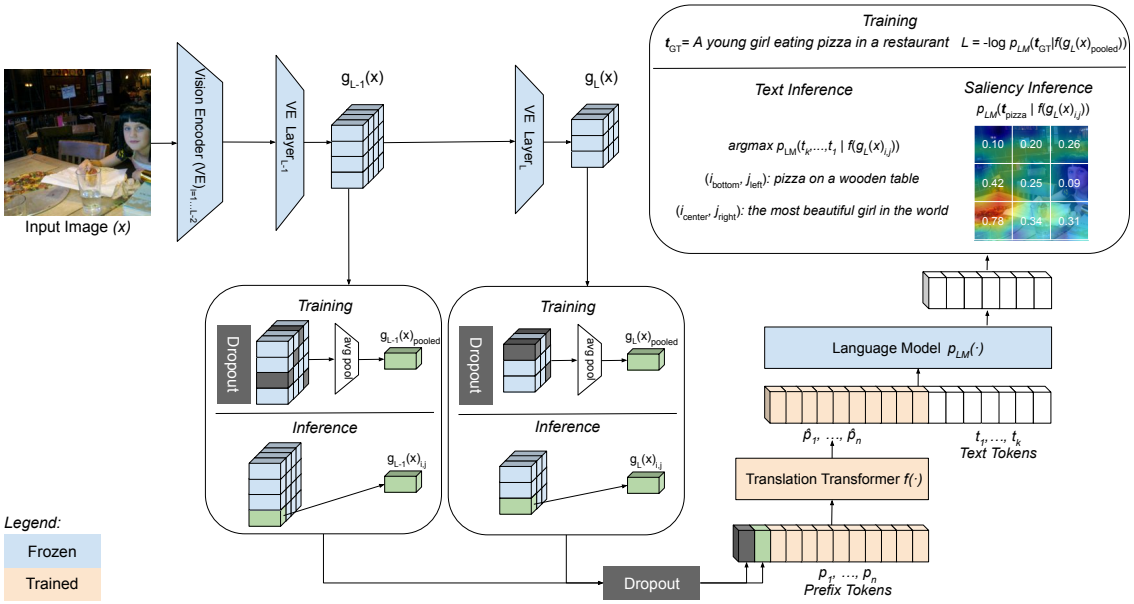


Figure 5.2: Architecture diagram of the proposed model, *DeViL*: Given an input image x and a pretrained vision encoder with multiple layers(L), we concatenate the average pooling embedding($g_L(x)_{pooled}$) to learnable prefix tokens (size=10; as considered by ClipCap [250]). A translation transformer f then projects prefix tokens $\hat{p}_1, \ldots, \hat{p}_n$ used to condition language model p_{LM} to get the captions for the entire layer. During Inference, we can select individual feature $g_L(x)_{i,j}$ for any arbitrary location i, j of the layer and generate its respective textual description, with saliency for any open-vocabulary word. The blocks in blue represent frozen weights while others are trained.

natural language description. Since *DeViL* is trained on top of features from a frozen vision model g, we view it as a *post-hoc* explanation method, that is able to reason about which concepts have been learned by the vision model retrospectively. Figure 5.2 depicts an overview of the network architecture of *DeViL*, which we describe in more detail in the following sub-sections.

5.3.2 Training Process

DeViL is designed to be a lightweight model such that it can be trained and applied to existing vision backbones without requiring a long training time. Instead of training a language model, we make use of a pretrained language model with frozen weights to generate the natural language descriptions for vision features. Specifically, the language model $p_{\text{LM}}(t_k|t_1,\ldots,t_{k-1})$ is able to predict the next token t from its tokenizer's vocabulary given a sequence of previous tokens as context. As shown in recent works [250], [252], the text generation of language models can be guided by conditioning the model on learned prefix tokens instead of fine-tuning its weights.

We follow this approach and train only a translation network f that takes image features $g_l(x)_{i,j}$ and produces n prefix tokens $\hat{p} = \{\hat{p}_i\}_{i=1}^n$ that are used to condition the language model $p_{\text{LM}}(t_k|\hat{p}_1,\ldots,\hat{p}_n,t_1,\ldots,t_{k-1})$ for generating a description. The translation network uses a standard Transformer architecture and takes as input the concatenated sequence of image features

$g_l(x)_{i,j}$ and the set of trained parameters for the initial prefix tokens p. From the output, we only keep the transformed prefix tokens $\hat{p}$ to pass on to the language model as prompt in a causal language generation setting.

Obtaining fine-grained data on textual descriptions for neurons or activations of the neural networks as done in [33] is costly at scale. Hence, we resort to more commonly available text-image pairs in the form of a captioning dataset to achieve the same goal. Since these image-text pairs relate the global image content to a single sentence, we describe in the following how we can train on these global relations, but evaluate our model on more fine-grained feature locations across the layers of the vision model.

Given a vision model g, we first specify the layers we would like to decode. For every forward pass of g, we extract the features $g_l(x)$ for every layer l in our set of explained layers. At training time, we perform 2D GAP over the spatial locations at each feature layer to obtain a single feature vector per layer $g_l(x)_{\text{pooled}}$. After applying a linear projection onto the same dimensionality and adding a positional embedding, the feature vectors are passed to the translation network f as a sequence. Depending on how many layers are being explained, the sequence length for f varies, i.e., it increases by one for each layer that is explained. At inference time, instead of pooling the visual features, we can select a specific location we would like to explain and only pass that particular feature vector $g_l(x)_{i,j}$ to f.

To train the translation network, we optimize the standard causal language modelling task of predicting the next token given previous tokens. In this context, the language model is conditioned on the image features through the translation network. Thus, the loss is

$$\mathscr{L} = \log p_{\text{LM}}(t|f(g(x))) \tag{5.1}$$

$$= \sum_k \log p_{\text{LM}}(t_k|\hat{p}_1,\ldots,\hat{p}_n,t_1,\ldots,t_{k-1}) \tag{5.2}$$

with trainable parameters only in the translation network f. Once trained, *DeViL* can decode vision features into natural language by conditioning the language model on a vision feature vector and generating a sentence. In practice, we greedily choose the most probable next word with arg max.

5.3.3 Dropout

The GAP at test time is required because we do not have more fine-grained data to train on. However, it creates a discrepancy between training and inference time. To overcome this issue, we introduce two dropout operations. Firstly, we randomly dropout spatial locations i, j of $g_l(x)$ before applying average pooling to obtain $g_l(x)_{\text{pooled}}$. As a result, our translation network observes a larger space of image features that better cover the full distribution of a vision layer's features. Secondly, we randomly subselect the layers from which the translation network f obtains the pooled features. This way, the translation network sometimes receives features only from individual layers during training time. This is crucial because, at inference time, we would like to decode a specific feature of an individual vision layer so this dropout ensures this input configuration is

seen during training. Moreover, it allows to train a single network per vision model instead of one per layer.

In contrast to dropout as proposed by [254], we do not remove features element-wise but always remove a full vector to simulate a spatial location or full layer being missing from the input. Hence, we also do not require to perform any rescaling of the features but rather mask the complete vectors from subsequent operations.

5.3.4 Open-vocabulary Saliency

DeViL can be used to obtain the probability of a given word or a phrase conditioned on vision features at layer l and location i, j by evaluating $p_{\text{LM}}(t_{\text{query}}|f(g_l(x)_{i,j}))$. When passing overall features from a layer of interest, we obtain a likelihood of the query for every spatial location that we can visualize as a saliency map. By using a general purpose language model, there are no constraints on the query such that we can obtain saliency maps on concepts that lie outside the original training scope of the vision model. This is useful for seeing whether the vision features encode information about related concepts due to their correlation with training data or obtained as a side-effect.

5.3.5 Experiments and Results

We evaluate *DeViL* on both its natural language and saliency generation capabilities. To train *DeViL* we use the CC3M [28] captioning dataset. While our goal is to generalize to more fine-grained descriptions of vision features, *DeViL* can still be used as a captioning model by itself. Therefore, we start by evaluating image captioning on CC3M, before discussing explanations of vision features obtained through *DeViL*. For fine-grained analysis, we report both qualitative as well as quantitative results on fine-grained neuron descriptions by training and evaluating on the MILANNOTATIONS dataset [33]. Lastly, we evaluate both textual descriptions and saliency obtained through *DeViL* across different layers to show its generalization capabilities. As we focus on explaining the vision backbone instead of producing captions, all images used for qualitative analysis come from sources outside CC3M, such as ImageNet [31], Places365 [32], and COCO [30]. Details about the *DeViL* architecture and training details can be found in Appendix B.

5.3.5.1 Image Captioning

Dataset We use the official train-validation split of the CC3M [28], consisting of 3M image-caption pairs collected from the web. We chose this dataset because it is sufficiently large to cover a large variety in both the vision and language modalities. Its successor CC12M [189] is less focused on high conceptual relevance, and more noisy in caption quality, making CC3M more suitable for tasks that require strong semantic alignment between text and image.

Baselines Although our goal is to translate latent representations of pretrained vision models into language, *DeViL* can still be used to obtain full image descriptions. Hence, we evaluate *DeViL* generated sentences with standard captioning metrics and compare against captioning methods [250], [252], [253]. ClipCap [250] is a lightweight model that combines the CLIP vision model with a pretrained language model. Similar to our approach, ClipCap only trains a translation network to keep the training cost low. Both UnifiedVLP [252] and LEMON [253] are large-scale models pretrained on several datasets not limited to captioning and subsequently fine-tuned. Thus, they surpass lightweight models such as ClipCap and *DeViL*, but require a lot of resources to train.

Results We present our results on common language-based metrics: BLEU-4 [222], METEOR [255], ROUGE-L [256], CIDEr [257], and SPICE [258]. We consider ResNet50 [89] trained on ImageNet and CLIP [235] in both its ResNet50 and ViT variants as our vision backbones.

We report image captioning results in Table 5.1. Between vision backbones, we observe that CLIP-ViT performs better than its ResNet50 counterpart and ResNet50 trained on ImageNet, which is not surprising given CLIP's contrastive vision-language pre-training. Since the vision encoder of CLIP has in the past depicted strong zero-shot capabilities on a variety of tasks, we would expect it also to have a large coverage of visual concepts when we explain their features.

When using the CLIP-ViT backbone, we further ablate the relevance of the pretrained language model. We make use of Open Pre-trained Transformers (OPT)-125M [259] and GPT-2 [184] in our pipeline. With CLIP-ViT-B-32 as the vision backbone and OPT as the language model, we obtain our best scores surpassing ClipCap on all by a big margin, e.g., a CIDEr of 86.29 vs. 71.82. Even when using GPT-2 and the same translation network architecture as ClipCap, we perform better across the board while using fewer parameters (40M vs. 43M). This is due to our model changes in using multiple layers of the vision backbone and the introduction of both feature-level and layer-level dropouts. Compared to large-scale captioning models that require more resources, we still perform better than UnifiedVLP on all metrics despite it requiring 1200h of training on a V100 GPU. In comparison, ClipCap reports 72h of training on GTX1080, while *DeViL* requires 24h of training on an A100. LEMON [253] still surpasses our lightweight model as the state-of-the-art model on captioning. Required training resources for LEMON are not reported.

Ablations We ablate the proposed token and feature dropouts in Table 5.2 in terms of captioning metrics for the CLIP-ViT model. This table shows 3 different models trained with incremental combinations of both dropouts. While each *DeViL* model is trained to explain multiple vision backbone layers, we evaluate them in two scenarios: using vision features from all layers of the vision backbone it was trained with (all) and when using only the last layer (single). This ablation study confirms that dropout is essential for producing image-relevant captions for individual layers at inference time. Since the model without token dropout has never seen a single layer being passed to the translation network, it performs poorly in this out-of-distribution scenario making CIDEr drop from 75.88 to 4.729. Feature dropout is less essential, but further improves captioning scores and reduces overfitting. When we compare results on using all layers rather than just using one,

Table 5.1: Image captioning results on CC3M [28] with different pretrained vision backbones and language models, and a comparison with state-of-the-art captioning models, either fully [252], [253] or partially-fine-tuned [250]. We report standard captioning metrics, where higher is better. IN-ResNet50: ImageNet [31] pretrained ResNet50 [89]. CLIP-ResNet50/ViT: ResNet50/ViT versions of the CLIP [235] vision encoder trained on the CLIP dataset.

Method	Vision Backbone	LM	B4 ↑	M ↑	RL ↑	C ↑	S ↑	#Params (M) ↓
DeViL	IN-ResNet50	OPT	5.851	9.572	23.92	65.73	15.23	88
	CLIP-ResNet50	OPT	6.770	10.68	25.90	78.41	17.38	88
		OPT	**7.506**	**11.22**	**26.82**	**86.29**	**18.37**	88
	CLIP-ViT	GPT-2	6.349	10.55	25.70	76.55	17.81	40
ClipCap [250]		GPT-2	-	-	25.12	71.82	16.07	43
VLP [252]			-	-	24.35	77.57	16.59	115
LEMON [253]			**10.1**	**12.1**	-	**104.4**	**19.0**	196.7

we see an improvement in all scores, e.g., CIDEr increases from 75.88 to 86.29. This suggests that complementary information is encoded in the different layers and caption content indeed differs, making it reasonable to assume we can obtain layer-specific explanations even when training on caption data.

Overall, these results show that with our dropout methods, we can train on several layers at once and perform an evaluation on individual layers by themselves, while also avoiding the need to train one model for each layer we might want to explain later on.

Table 5.2: Ablation study of the proposed dropout mechanism when evaluated using vision features from all layers or only the very last layer (single) of CLIP-ViT.

Layer	Token Dropout	Feature Dropout	B4 ↑	M ↑	RL ↑	C ↑	S ↑
single			0.3038	3.435	11.76	4.729	1.962
	✓		6.442	10.34	25.14	73.73	16.86
	✓	✓	6.623	10.49	25.50	75.88	17.11
all			7.101	11.16	26.62	82.59	18.54
	✓		7.211	11.09	26.70	83.32	**18.48**
	✓	✓	**7.506**	**11.22**	**26.82**	**86.29**	18.37

5.3.5.2 MILAN

Since *DeViL* was designed first and foremost for localised feature inspection in natural language, we strive to compare our method more directly on an explainability task, and especially on one such task at the local feature level. Thus, we also train *DeViL* on MILANNOTATIONS [33], a dataset of over 50k human descriptions of sets of image regions obtained from the top neuron activating images of several known models like ResNet [89] or AlexNet [260]. For each base

model, the authors collect descriptions for the top 15 activated images of each neuron in the network. Each image is masked by its corresponding activation mask, highlighting only the regions for which the corresponding neuron fired.

We compare with ClipCap and the MILAN model [33] trained on MILANNOTATIONS [33]. Although our model was designed to work with layer-wise feature maps for a single image and not at the neuron level, we adapt *DeViL* by pooling over the 15 masked images given for each neuron. We report the NLP metrics including BERTScore [211] results in Table 5.3. The results are averaged over several generalization experiments proposed by [33]. We report a complete comparison in terms of BERTScore of all 13 experiments in Appendix B.

The average BERTScore across scenarios is 0.362, 0.381 and 0.382 for MILAN, ClipCap and *DeViL*, respectively. The margins between all models are small as BERTScore is very sensitive to small differences that can still indicate a reliable ranking [238]. Considering all other language metrics, the difference between ClipCap and *DeViL* is more pronounced and *DeViL* outperforms ClipCap consistently. A qualitative comparison with MILAN [33] can be seen in Figure 5.3. We refer the reader to examples like "animals and chains" and "building and windows". Both quantitative and qualitative results validate *DeViL's* generalization ability and its primary intended goal: to faithfully decode localised vision features into natural language.

Table 5.3: Evaluation of MILAN, ClipCap and *DeViL* on MILANNOTATIONS.

Method	B4 ↑	M ↑	RL ↑	C ↑	S ↑	BS ↑
MILAN [33]	-	-	-	-	-	0.362
ClipCap [250]	3.99	9.62	27.0	25.1	10.8	0.381
DeViL	**6.28**	**11.3**	**30.6**	**33.7**	**13.3**	**0.382**

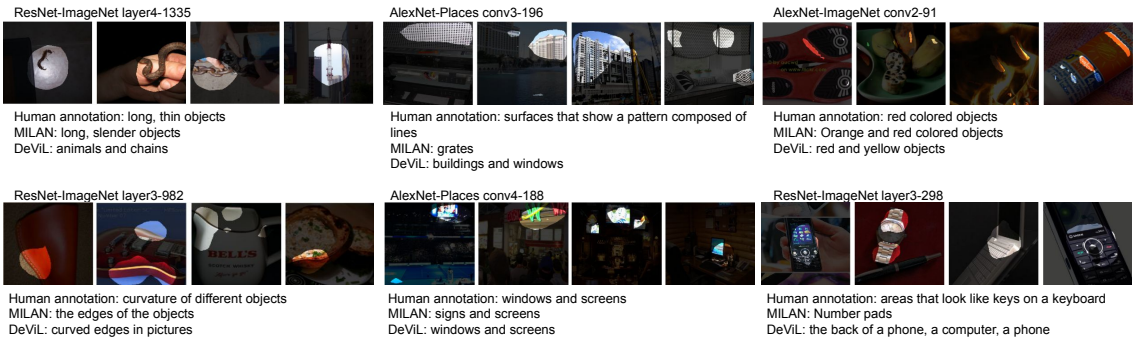


Figure 5.3: Qualitative results on MILANNOTATIONS [33] and comparison with the MILAN [33] model.

5.3.5.3 Layer-wise Explanations

DNNs learn to extract meaningful patterns from input data by progressively building up a hierarchy of features. The lower layers tend to detect simple patterns like edges and curves, while the higher layers learn to recognize more complex and abstract concepts [89], [237]–[239]. We verify a

similar trend in the descriptions generated by *DeViL*. We generate descriptions for each spatial location of the feature map at layer l and produce saliency maps to measure the spatial support of the main subject in the sentence generated at a single location. Our model assigns a probability score to a textual query being generated based on its relevance to the visual features. This score can then be visualised as a heatmap, with higher scores corresponding to areas of the image where the vision model encoded the textual concept.

5.3.5.4 Qualitative Results

Figure 5.4 corresponds to the output generated by the CLIP-ResNet50 model for layers 21, 39, and 49. The green dot in the image shows the location for which the description has been generated. The generated descriptions showcase the aforementioned hierarchy, with lower layers mentioning lower level features like colours (e.g., "red headlight") and shapes (e.g., "squares with a circle") and higher layers mentioning objects (e.g., progression from "building" to "church"). The saliency maps also validate the spatial location of these words, e.g., higher saliency scores for "zebra" at L49, "animal" at L39, and "pattern" at L21.

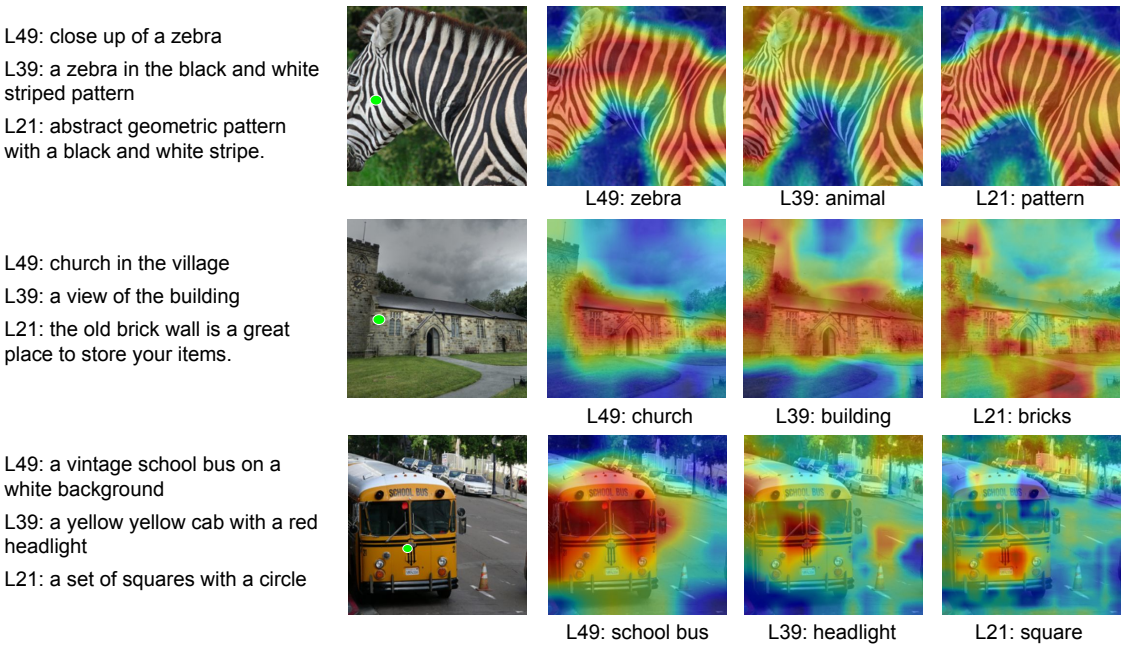


Figure 5.4: Captions generated for each image by the CLIP-ResNet50 model, for layer 49 (L49), layer 39 (L39) and layer 21 (L21). To the right of original image, we present saliency maps for relevant words taken from the captions at the location marked with a green dot. With each layer the learned concepts go from higher level (more specific) to lower level (more general).

We also show textual descriptions generated by different layers of the CLIP-ViT model. A key feature about ViTs is that they tend to generate similar feature maps in between blocks, which can be attributed to the way the self-attention mechanism works [47], [261]. For this reason, we evaluate the last 12 layers of transformer-based models in contrast to 3 layers in convolution-based models. As shown in the school bus image of Figure 5.5, CLIP-ViT generates the word "bus" in

its last layer (L12), and moves to lower level concepts including the identification of the colour "yellow" for L6, and lastly the shape "square" at L3. We try to capture a wide variety of objects in our qualitative analysis to show the generalization capability of the model.

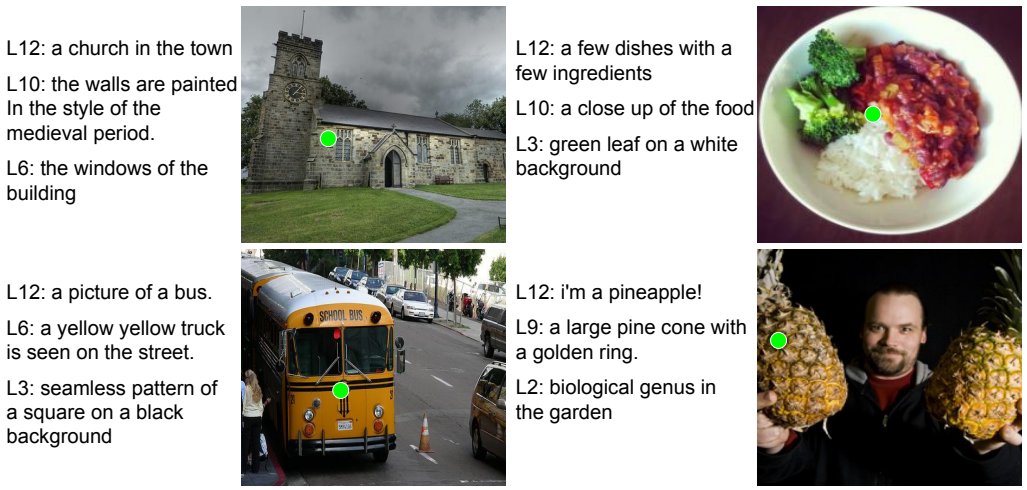


Figure 5.5: Descriptions generated for CLIP-ViT at different layers for the locations indicated by the green visual marker.

5.3.5.5 Layerwise Descriptions

To quantify how well *DeViL* captures the differences of what specific vision layers encode, we analyze the generated text descriptions across layers. Specifically, we obtain a single text description per layer by averaging the vision features for each layer and decode them individually with *DeViL*. In Table 5.4, we take a look at the properties of the language generated by *DeViL* on CC3M. As measured by the CIDEr score, we observe that the text similarity to human captions increases with later layers as more semantically meaningful embeddings are produced by the vision model. We also see that the language shifts from using many adjectives in earlier layers to more verbs and a more comprehensive vocabulary in higher layers. This indicates a shift from describing simple concepts with a limited set of attributes to richer descriptions that also describe relations and actions between objects in the image. In Figure 5.6, we show the saliency for the same word across different layers of CLIP-ResNet50. The top layer encodes "Zebra", but not "Pattern", and conversely for a lower layer. These insights can help us better understand intermediate features of a pretrained model and assist in choosing the right features for a downstream task.

5.3.5.6 Spatial and Layer Coherence

We want to further analyze how well the spatially localised descriptions generated by *DeViL* capture the content of the underlying image patches. Intuitively, earlier layers encode local information of a smaller underlying patch, while later layers can encode more global information. To test this, we create center crops of sizes 32x32, 64x64, and 128x128 along the full image size of 224x224 and compare their similarity to *DeViL* descriptions of the vision feature closest to the

Table 5.4: Per-layer statistics of adjectives, verbs, and number of unique words.

Model	Layer	Adj (%)	Verbs (%)	#Uniq. Words	CIDEr (↑)
CLIP-ViT	9	6.37	6.30	4790	61.6
	6	7.60	5.78	4115	44.1
	3	8.04	5.24	2554	20.2
CLIP-ResNet50	49	6.30	6.66	5228	71.2
	39	6.69	6.43	4395	54.1
	21	6.89	5.90	3374	34.6

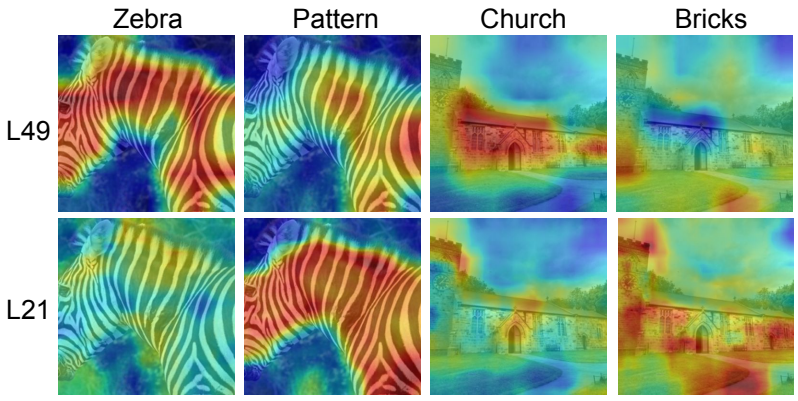


Figure 5.6: Examples of open-vocabulary saliency maps showcasing that semantics are encoded in different layers (CLIP-ResNet50 model).

center in each layer's feature map. We use CLIP-ResNet50 to embed the cropped images as well as *DeViL* descriptions and compare their similarities. Figure 5.7 plots which crop size is most similar with the generated text. We observe that smaller patches have much higher similarity with earlier layer descriptions, and conversely bigger patches with later layers. Descriptions of L21 best match 64x64 patches and this distribution shift progressively to L49 fitting best to the full image content. These results validate that *DeViL* exposes what the vision model encodes both in lower layers describing the content of local patches and global concepts in higher layers.

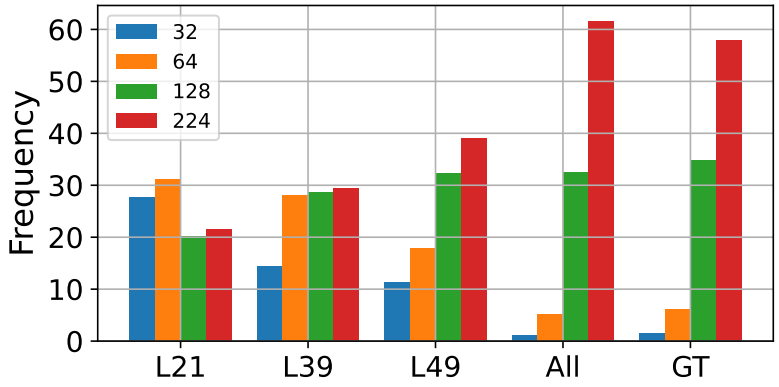


Figure 5.7: Frequency of most similar patch size for a given *DeViL* description (L21, L39, L49, All) and human caption (GT) using CLIP-ResNet50.

5.3.5.7 Inspecting Different Vision Models

Since *DeViL* incorporates an LLM, it has the ability to produce saliency maps for any word, not being limited by the closed-set of classes of a given dataset. Figure 5.8 presents a comparison of different vision backbones in terms of saliency maps for different words. The different models can distinguish "cat" and "dog", but also identify both as "animal". Interestingly, since CLIP-based models are trained with not only images but also text, they can identify the written word "pizza" on the bottom-most image, while an ImageNet trained ResNet50 cannot. Similarly, the word "pirate", written on a stone, is also identified by CLIP-based models for the aquarium image. These failure cases of the CLIP model have been discovered independently and the explanations of our *DeViL* model expose these as well, highlighting its faithfulness to the vision model. Obtaining saliency maps for open-vocabulary words that did not appear in a model's training data such as the supercategory "animal" or the class "pirate" is a novel contribution of our method. It also allows a more direct comparison with models of different pretraining tasks such as CLIP, which is not possible with methods that rely on the model's output to be the same for direct comparisons. More qualitative results can be found in Appendix B.

We also compare our saliency maps quantitatively with Grad-CAM [5], which proposes interpreting a neural network via the gradients for a target class flowing through a given layer. Results are presented in Table 5.5 for 3 layers of the ResNet50 model (L49, L39, and L11). We report commonly used deletion (Del.) and insertion (Ins.) metrics [103]. The Del.(Ins.) metric measures the drop(rise) in the probability of a class as relevant pixels given by the saliency map are gradually removed(inserted) from the image. Thus, lower(higher) is better for Del.(Ins.). Although we are unable to outperform Grad-CAM, the latter uses gradients, which provide more information than just activations alone. Furthermore, Grad-CAM is specialised in the network's output classes, while *DeViL* can perform open-vocabulary saliency. An interesting future research direction is to improve and align open-vocabulary saliency with existing approaches.

Table 5.5: Comparison between the ResNet50 model and Grad-CAM [5] for saliency maps generated for different layers (L49, L39, and L21), in terms of deletion (↓) and insertion (↑) metrics.

Layer	Metric	Grad-CAM [5]	*DeViL* (CC3M)
L49	Del.	0.342	0.466
	Ins.	0.763	0.697
L39	Del.	0.465	0.494
	Ins.	0.677	0.661
L21	Del.	0.405	0.484
	Ins.	0.593	0.635

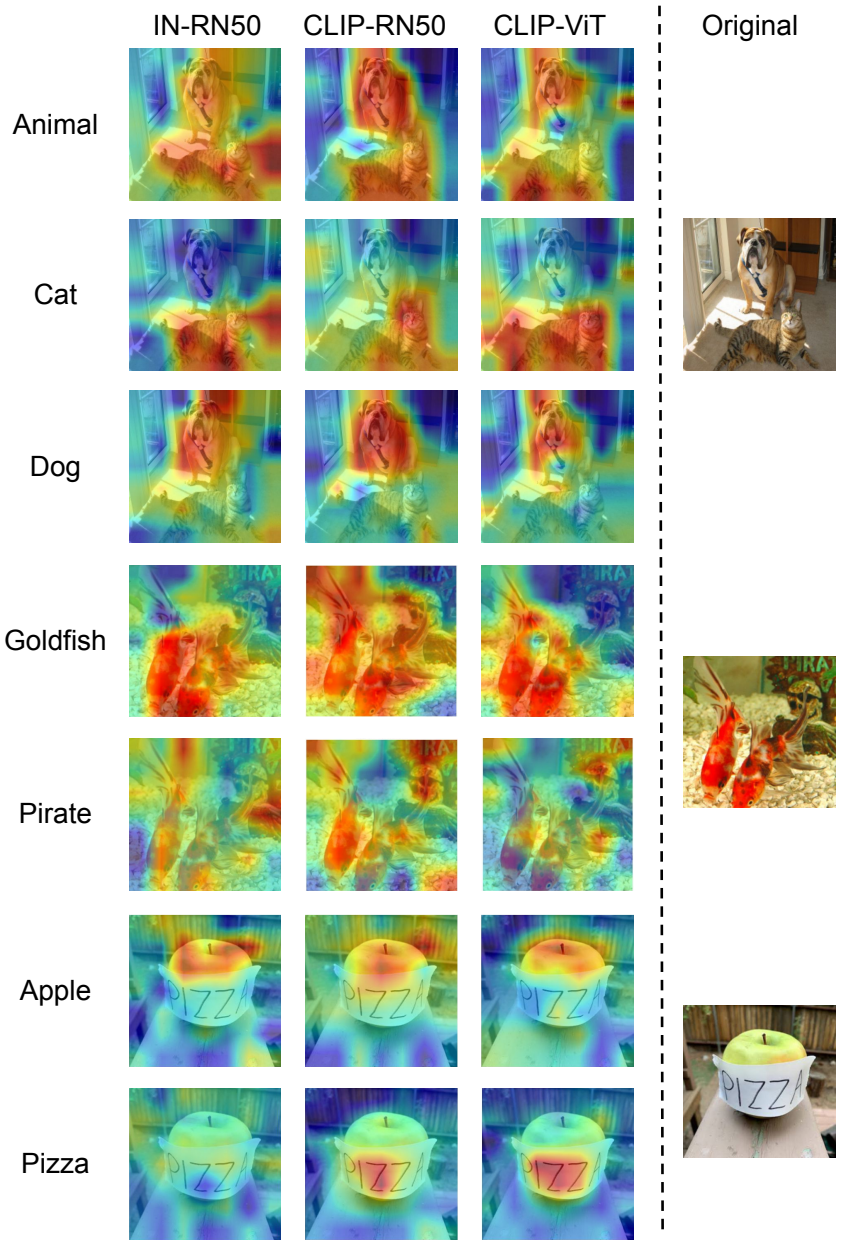


Figure 5.8: Open-vocabulary saliency maps for different backbones with the final layer. For CLIP-ViT we use the penultimate layer.

5.3.6 Summary and Conclusions

We proposed *DeViL*, a unique approach for explaining feature maps of any vision encoder through natural language, without requiring task-specific heads. Besides generating textual descriptions for individual features at different layers, it can also produce open-vocabulary saliency attributions. We evaluate the efficacy of *DeViL*'s language generations on CC3M and MILANNOTATIONS, outperforming competing models, and show extensive qualitative results validating that *DeViL*'s open-vocabulary saliency exposes which concepts are understood by each of the model's layers.

5.4 Understanding Chest X-ray Vision Representations with *DeViL*

Having introduced *DeViL* and validated it in multiple natural image analysis tasks, we are now interested in assessing if and how *DeViL* is useful for understanding the representations learned by CXR vision encoders. We adapt *DeViL* for this purpose by training on structured radiology report data and by leveraging a radiology-specialised language model. In the end, the resulting model aligns visual features with clinically meaningful concepts while preserving the flexibility of the original approach. We conduct experiments on Structured Radiology Report Generation (SRRG), saliency generation, and open-vocabulary saliency-text grounding on different types of chest X-ray vision encoders: convolutional, self-supervised ViT, and vision–language Transformer. Results show competitive performance with large end-to-end report generation models and demonstrate that *DeViL*'s open-vocabulary saliency maps outperform those produced by a specialised saliency generation method for vision-language encoders.

5.4.1 Methodology

While the original *DeViL* is trained on natural image captioning datasets, where each image is paired with a single sentence describing the whole image, radiology reports describe multiple findings within the X-ray, referring to different anatomical regions and clinical observations. Although such reports provide more detailed supervision, they also introduce additional complexity in aligning with visual features, due to their high variability (e.g., variable number of sentences per report, differences between reports for healthy vs diseased cases). To be able to leverage the additional supervision offered by radiology reports without increasing training complexity, we train *DeViL* on the recently introduced task of SRRG [34], where unstructured reports from the MIMIC-CXR [13] and CheXpert Plus [193] datasets have been decomposed into standardised sections describing individual findings or anatomical regions.

Additionally, instead of the general-purpose LLM used in *DeViL*, in this work we employ RadPhi-2 [262]. Using a radiology-specialised language model simplifies training because the Translation Transformer and the prefix tokens learned to condition the LLM can focus primarily on aligning visual features with clinical concepts. Since the language model already captures the distribution of clinical terminology and reporting style, *DeViL*'s trainable components do not need to learn a transformation from general natural language to medical language. Instead, they only need to steer an already clinically grounded language distribution, reducing variability in the learning process and simplifying the vision–language alignment task.

5.4.2 Experiments and Results

5.4.2.1 Implementation Details

We train three versions of *DeViL* to be able to explain and compare three different types of vision encoders:

1. a CNN: DenseNet-121 [88] from TorchXRayVision[1] [263], trained on the MIMIC-CXR dataset [13].

2. a ViT trained with self-supervised representation learning: XrayDINOv2[2] [193]. DINOv2 learns visual representations using a teacher–student self-distillation framework, encouraging the model to produce consistent representations for different augmented views of the same image. The X-ray adapted version is trained on multiple large-scale CXR datasets, namely MIMIC-CXR [13], CheXpert Plus [193], PadChest [194], BIMCV [195], and Candid-PTX [264].

3. a Transformer-based VLM based on CLIP [235], but instead of using contrastive learning, it is trained with Sigmoid Loss for Language-Image Pretraining (SigLIP) [236]. Unlike purely visual encoders, SigLIP learns joint image–text representations by aligning images with associated textual descriptions. The X-ray adapted model, XraySigLIP[3] [193], is trained on the same data as XrayDINOv2.

All versions are trained on the SRRG-Impression dataset [34], derived from the MIMIC-CXR [13] and CheXpert Plus [193] datasets. Training was carried out on an A100 80GB GPU for 500000 steps, with a batch size of 16, and a linearly decaying learning rate of 1×10^{-4}. We keep the same parameters of the original *DeViL* [29], i.e., 10 learnable prefix tokens, 12 layers (hidden size of 3,072, 12 attention heads) for the Translation Transformer, and 0.5 probability for both token/layer and spatial feature dropout.

5.4.2.2 Structured Radiology Report Generation

We start by evaluating *DeViL* on the task it was trained on, i.e., SRRG, and compare it against the SRRG-fine-tuned models introduced by Delbrouck *et al.* [34]:

1. CheXagent [115], a foundation VLM for CXR interpretation trained in three stages: clinical text pretraining of the LLM, SigLIP-based [236] image–text alignment for the image encoder, and joint multimodal instruction tuning for diverse CXR tasks.

2. CheXpert-Plus [193], composed of a SwinV2 ViT [265] as image encoder and a lightweight two-layer Transformer decoder for SRRG.

3. MAIRA-2 [266], a multimodal autoregressive model for grounded CXR report generation, with a LLaVA-based architecture [120], combining Rad-DINO [267] as image encoder, Vicuna-7B-based LLM [268], and a trainable adapter.

Table 5.6 reports the results using standard text generation metrics (BLEU [222], ROUGE-L [256], BERTScore [211]) as well as clinically oriented metrics from the SRRG benchmark:

[1] TorchXRayVision checkpoint: `densenet121-res224-mimic_ch`
[2] HuggingFace checkpoint: `StanfordAIMI/dinov2-base-xray-224`
[3] HuggingFace checkpoint: `StanfordAIMI/XraySigLIP__vit-b-16-siglip-512__webli`

F1-RadGraph [269] and F1-SRR-BERT (Precision, Recall, and F1-Score) [34]. F1-RadGraph evaluates the overlap between the clinical entities and their relations extracted from the ground-truth and generated reports. F1-SRR-BERT computes the F1-score between the 55 hierarchical SRR-BERT-predicted radiological labels of the ground-truth and generated reports.

Among the three *DeViL* variants, the SigLIP backbone achieves, as expected, the best overall performance in traditional metrics (ROUGE-L 28.71, F1-RadGraph 19.08), while the DINOv2 version obtains the highest F1-SRR-BERT of 42.79. Compared to the baselines, *DeViL* achieves the highest ROUGE-L, F1-RadGraph, and F1-SRR-BERT Precision values. Overall, all *DeViL* models achieve competitive clinical F1 scores compared to large foundation models fine-tuned for SRRG, despite keeping both the vision encoder and the language model frozen, and despite SRRG not being the main goal of *DeViL*.

Table 5.6: Performance on the SRRG-Impression dataset. Best results are in bold, second-best are underlined.

Model	Traditional Metrics				F1-SRR-BERT		
	BLEU	ROUGE-L	BERTScore	F1-RadGraph	Precision	Recall	F1-Score
CheXagent	4.68	26.10	<u>59.70</u>	18.33	<u>45.24</u>	<u>56.70</u>	**48.64**
CheXpert-Plus	**14.07**	26.79	59.21	<u>18.89</u>	43.46	48.15	44.56
MAIRA-2	5.28	26.61	**60.79**	**19.08**	44.80	**57.69**	<u>47.97</u>
DeViL DenseNet	5.907	25.89	55.58	17.01	43.28	40.83	38.14
DeViL XRayDINOv2	6.961	<u>28.01</u>	57.75	18.88	**45.36**	47.32	42.79
DeViL XRaySigLIP	<u>7.098</u>	**28.71**	57.72	**19.08**	44.89	46.34	42.07

5.4.2.3 Visual Saliency for Classification

Table 5.7: Assessment of *DeViL* vs. Grad-CAM saliency map quality on Faithfulness and Localisation metrics. # inst. is the number of generated saliency maps (i.e., the number of positive predictions of each vision backbone). # masks is the number of instances used for the computation of the Att. Loc. and Rel. Rank Acc. metrics (i.e., the number of aforementioned instances for which there are ground-truth segmentation masks available). Results where *DeViL* outperforms GCAM are in bold.

Model	XAI Method	AUC	# inst.	Faithfulness				Localisation			
				Deletion ↓	Insertion ↑	Conf. Drop ↓	Conf. Incr. ↑	Focus ↑	# masks	Att. Loc. ↑	Rel. Rank Acc. ↑
DenseNet	Grad-CAM	70.36	1851	0.3897	0.5618	0.2272	0.0962	0.3340	127	0.1776	0.2168
	DeViL			0.4775	0.4805	**0.2269**	0.0616	0.2555		0.1296	0.1320
XRayDINOv2	Grad-CAM	76.77	1548	0.5858	0.7282	0.1363	0.4780	0.3178	115	0.2301	0.2708
	DeViL			0.6658	0.6928	**0.0784**	**0.4890**	0.2524		0.1149	0.1441
XRaySigLIP	Grad-CAM	80.41	1472	0.4214	0.6985	0.1500	0.3193	0.3370	122	0.2867	0.3542
	DeViL			0.5527	0.6382	**0.0551**	**0.5428**	0.2526		0.1241	0.1607

We start by comparing the quality of *DeViL*'s saliency maps against Grad-CAM's [5]. Although *DeViL* explains the vision encoder without needing a task-specific head and the corresponding prediction, Grad-CAM requires it, so we train a multilabel classification layer (linear probe) on top of the frozen vision encoders on the SRR-Impression dataset [34]. Training is done

Table 5.8: Assessment of *DeViL* vs. NIB saliency map quality on Faithfulness and Localisation Metrics for the XRaySigLIP vision encoder. Best results in bold.

XAI Method	Faithfulness				Localisation		
	Deletion ↓	Insertion ↑	Conf. Drop ↓	Conf. Incr. ↑	Focus ↑	Att. Loc. ↑	Rel. Rank Acc. ↑
NIB	0.6450	0.5256	0.0895	0.3573	0.2499	0.1104	0.0748
DeViL	**0.5527**	**0.6382**	**0.0551**	**0.5428**	**0.2526**	**0.1241**	**0.1607**

for 10 epochs, with a learning rate of 1×10^{-3} and a batch size of 128. For saliency map quality assessment, we generate maps for all positive prediction of all images in the test set; for *DeViL*-based models we use the name of the class to generate the saliency maps.

Table 5.7 presents these results. Besides different saliency map quality metrics, we report the AUC obtained by the aforementioned linear probes and the total number of generated explanations (i.e., the total number of positive predictions). As anticipated, DenseNet achieves the lowest AUC, since the vision encoder has been trained only on MIMIC-CXR [13], while XRayDINOv2 and XRaySigLIP have been trained on several datasets, including both MIMIC-CXR [13] and CheXpert Plus [193] (i.e., the datasets that compose the SRRG-Impression dataset [34]). XRaySigLIP achieves the highest AUC of 80.41, demonstrating clear benefits from Vision-Language (VL) pretraining.

Saliency quality is evaluated with the Quantus framework [101] across two dimensions: faithfulness (whether explanations reflect the model's predictions) and localisation (whether saliency falls within ground-truth regions of interest). Faithfulness is measured using: i) Deletion/Insertion metrics [103], [104], which track the drop/rise in class probability as image patches are progressively removed/inserted according to saliency-assigned relevance; and ii) Confidence Drop/Increase metrics [248], which measure how model confidence changes when the input is multiplied by the class saliency map. A good saliency map highlights regions most relevant to the decision, so keeping only those regions should cause a small confidence drop. Confidence Drop measures the difference between the model's confidence on the original image and on the saliency-masked image, while Confidence Increase counts how often confidence rises when irrelevant regions are removed (reported as a percentage over the test set).

Localisation is assessed using: i) the Focus metric [109], which quantifies saliency precision by creating a four-image mosaic where only one image belongs to the explained class and should receive positive attribution; ii) Attribution Localisation [110], which computes the ratio of positive attributions inside the target object relative to all positive attributions; and iii) Relevance Rank Accuracy [270], which measures the proportion of highly attributed pixels within a ground-truth mask. For the last two metrics, we use the radiologist-annotated CheXlocalize dataset [36], evaluating all positively predicted diseases for test images with available segmentation masks (see column '# masks' in Table 5.7).

In terms of faithfulness, Grad-CAM XRayDINOv2 achieves the best overall result, except in the Deletion metric, where it ranks last. However, in terms of localisation, Grad-CAM XRaySigLIP outperforms the other two, once more corroborating the hypothesis that VLpretraining not only

improves classification performance, but also leads to more focused explanations. Although *DeViL* does not surpass Grad-CAM, especially in localisation, it shows competitive results in both faithfulness and localisation metrics. This is expected since Grad-CAM uses gradient information from the linear probe, while *DeViL* does not. However, the XRayDINOv2 and XRaySigLIP variants significantly surpass their Grad-CAM counterparts in confidence drop and increase. Finally, a trend emerges amongst the different *DeViL* variants: XRaySigLIP surpasses XRayDINOv2, which, in turn, surpasses DenseNet, where confidence drop, confidence increase, and relevance rank accuracy are concerned. This again points towards the higher effectiveness of self-supervised (XRayDINOv2) and VL(XRaySigLIP) pretraining.

While Grad-CAM benefits from access to gradients of a task-specific classifier, which are more informative than attributions alone [29], *DeViL* operates directly on frozen encoders and still achieves competitive faithfulness and localization. Nevertheless, to ensure a fair and thorough evaluation, we also compare *DeViL* against a more closely related visual saliency method, NIB [234]. NIB is a *post-hoc* saliency method for multimodal models like CLIP [235], that measures each feature's contribution by progressively narrowing an information bottleneck and integrating how the model's image–text mutual information changes as features are scaled from fully active to suppressed. Since NIB can only be used on multimodal models, we apply it to the XRaySigLIP backbone and compare it with its corresponding *DeViL* version in Table 5.8. Even though both NIB and *DeViL* do not need classification heads to explain the underlying vision backbone, in this experiment we use the same experimental setting as in Table 5.7 (i.e., explaining all positive predictions given by the corresponding linear probes) for a direct comparison, and, as expected, similarly to *DeViL*, NIB underperforms when compared to Grad-CAM. *DeViL* clearly outperforms NIB in all faithfulness and localisation metrics, even though NIB is designed to specifically explain CLIP-based models, while *DeViL* can be applied to any vision encoder.

5.4.2.4 Open-Vocabulary Visual Saliency

One of the novelties introduced by *DeViL* [29] is its ability to generate open-vocabulary saliency maps given any word or sentence. To test if the generated saliency maps are well aligned with the target sentence, we use the MS-CXR dataset [37], which provides bounding boxes and their corresponding phrases. We test two scenarios: the first where the text label associated with the bounding box is given as the target for the saliency generation, and the second where the target is the entire sentence. Results are presented in Table 5.9, for all versions of *DeViL* and for NIB. We measure the localisation performance of the generated maps using Attribution Localisation and Relevance Rank Accuracy.

All *DeViL* versions overperform NIB, except DenseNet in the second scenario for the Relevance Rank Accuracy metric. When comparing the same vision backbone, XRaySigLIP, *DeViL* surpasses NIB in both scenarios. Finally, excluding NIB, metrics are slightly higher in the first scenario when comparing the same vision backbone, which indicates that *DeViL*'s maps are slightly better aligned with single words than entire sentences. Among the three backbones, XRaySigLIP

largely outperforms the other two in terms of Relevance Rank Accuracy, again in alignment with previous results.

Table 5.9: Open-vocabulary visual saliency evaluation on the MS-CXR dataset. 'Label'/'Sentence' corresponds to the scenario in which the methods are given an image, the bounding box, and its corresponding label/sentence. Best results in bold.

Text Input	XAI Method	Model	Att. Loc. ↑	Rel. Rank Acc. ↑
Label	*DeViL*	DenseNet	0.1243	0.1251
		XRayDINOv2	**0.1250**	0.1385
		XRaySigLIP	0.1216	**0.1514**
	NIB	XRaySigLIP	0.1144	0.0862
Sentence	*DeViL*	DenseNet	**0.1232**	0.1069
		XRayDINOv2	0.1190	0.1148
		XRaySigLIP	0.1200	**0.1551**
	NIB	XRaySigLIP	0.1157	0.1132

The same can be seen in Figure 5.9, which shows some open-vocabulary saliency maps produced by *DeViL* for the three backbones. Although DenseNet maps are smoother than maps from the other two methods (which is caused by the conversion of images into patches that is done in ViT-based models like XRayDINOv2 and XRaySigLIP), the maps produced for the XRaySigLIP backbone are more aligned with the ground-truth sentences (e.g., on the first row, the XRaySigLIP map is well circumscribed to the region of the bounding box, on the second row the upper part of both lungs has higher saliency, or in the last row where higher saliencies are concentrated to the left lung). However, for more global clinical concepts, such as the enlargement of the cardiac silhouette (see the third row) *DeViL* produces broader saliency maps. This happens with all vision encoders, which might suggest that the encoders themselves have difficulty in encoding less localised concepts.

Both quantitative and qualitative results support the trends observed in previous sections, suggesting that representations learned through self-supervised and VLpretraining yield more informative and clinically meaningful explanations.

5.4.2.5 Region-to-Text Generation

Through spatial feature dropout *DeViL* is able to generate descriptions of delimited regions of the input image. To test this ability, which is usually only achieved by highly specialised models like GRiT [271], we again use the MS-CXR dataset [37] and instead of pooling the vision features of the entire image, we pool only the features inside the ground-truth bounding box. Results are presented in Table 5.10, where the text descriptions generated with and without bounding box pooling are compared against the ground-truth sentence associated with the bounding box. The metrics show that there is a mismatch between the ground-truth and the *DeViL*-generated sentences, which is to be expected since *DeViL* was trained to generate structured radiology reports and not single, isolated sentences like the ones in MS-CXR. Nonetheless, when pooling only the

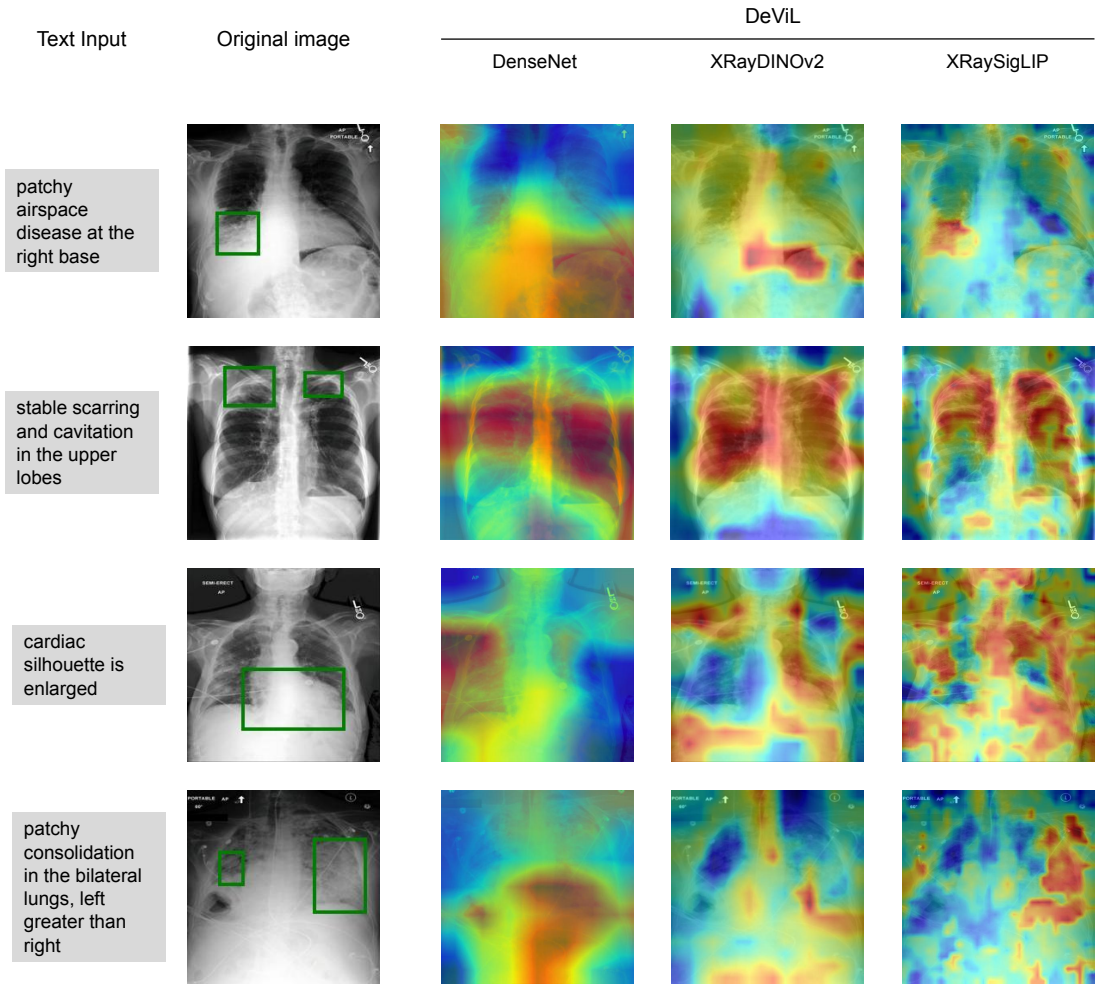


Figure 5.9: Qualitative examples of open-vocabulary saliency maps generated by the three *DeViL* versions for different images and sentences of the MS-CXR dataset. Bounding boxes indicate the region each sentence refers to.

image features inside the bounding box, the results improve, which shows that *DeViL* is indeed able to perform region-to-text generation.

Table 5.10: Region-to-text generation results of *DeViL* XRaySigLIP on the MS-CXR dataset with (✓) and without (✗) a bounding box. Best results in bold.

BBox	BLEU	ROUGE-L	BERTScore	F1-RadGraph
✗	0.4693	10.09	25.14	6.188
✓	**0.5076**	**10.29**	**27.50**	**6.854**

5.4.3 Summary and Conclusions

We adapted *DeViL* for understanding CXR models by using a radiology-specialised language model and training on structured radiology reports. This setup allows visual features to be translated into clinically meaningful concepts while preserving the flexibility of the original approach. Across structured report generation, saliency evaluation, and saliency–text grounding experiments, we analysed convolutional, self-supervised ViT and VL CXR encoders. Results show competitive performance in structured report and saliency generation, and demonstrate that *DeViL*'s open-vocabulary saliency maps surpass those of a VLM-specific saliency method, while remaining applicable to any vision backbone. These findings highlight the potential of language-based interpretability approaches for analysing medical vision representations and show that *DeViL* is a flexible framework for studying how different CXR models encode clinically relevant information, which can be useful for model auditing.

Chapter 6

Summary and Conclusions

The black-box nature of DL models hinders their deployment in clinical workflows. While current approaches (most often *post-hoc* visual explanations such as saliency maps) try to increase the interpretability of such models, they face critical limitations: they do not clarify why features matter, are often hard to interpret, and may not faithfully reflect the model's reasoning. Moreover, users with different expertise (e.g., clinicians vs. patients) require diverse forms of explanations, highlighting the need for versatile systems that combine multiple explanation modalities. However, generating some types of explanation modalities, especially concept-based explanations or NLEs, often requires costly clinician-annotated data, which can reduce both scalability and faithfulness.

In view of the challenges discussed, this thesis set out two primary objectives: first, to investigate and design alternatives to *post-hoc* visual explanations for medical image classification; and second, to achieve this without imposing additional annotation requirements, by employing various forms of weakly supervised learning.

With these objectives in mind, several novel and meaningful contributions were made, as introduced in Chapter 1 and described throughout this document:

- Chapter 2: we studied the application of visual explanations in medical image classification. First, through experiments on DR, skin lesion, and CXR classification, we showed that while attention mechanisms (whether explicitly incorporated into non-attention models or inherent to ViTs) are widely used for their supposed improved performance, this is not always the case. Moreover, there are also no guarantees that these mechanisms improve the interpretability (given by *post-hoc* saliency maps) of the underlying networks. Similarly, our study on B-cos networks, the first of its kind in the medical domain, revealed that although this inherently interpretable architecture generates more localised explanations through weight-input alignment introduced during training, its explanation quality advantages observed on ImageNet-trained models do not consistently transfer to medical tasks, and are accompanied by a slight reduction in classification performance.

- Chapter 3: we explored a different type of explanation modality, concept-based explanations. In this part of the work, we transformed LVLMs into CBMs. The proposed method-

ology, *CBVLM*, leverages the few-shot and training-free capabilities of LVLMs , thus operating under a weakly supervised paradigm that requires only a handful of annotated examples. By grounding predictions on concepts, *CBVLM* preserves the interpretability of CBMs while overcoming their limitations, namely the heavy annotation burden and the need for retraining when introducing new concepts. Extensive experiments across four medical datasets and a wide range of LVLMs demonstrated that *CBVLM* consistently outperformed both traditional CBMs and task-specific black-box models.

- Chapter 4: we focused on research into NLEs, by first addressing the fully supervised setting to demonstrate the feasibility of supervising NLE generation at the Encoder and sentence level within an Encoder–Decoder framework. We empirically validated the viability of Encoder-level supervision as a step towards NLE generation when GT explanations are scarce. Building on these insights, we introduced *WeNLEX*, a weakly supervised framework for generating faithful and clinically plausible NLEs via image feature reconstruction and distribution matching. *WeNLEX* can operate in both *post-hoc* and in-model settings, with the latter improving classification performance, highlighting the complementarity of explainability and predictive accuracy. Furthermore, *WeNLEX* has the ability to adapt explanations for different audiences, from clinicians to lay users, which underscores its practical relevance for real-world medical applications.

- Chapter 5: we investigated vision–language explanations through the introduction of *DeViL*, a framework that aligns the visual features of any vision encoder with natural language by training a lightweight translation network on image–caption pairs, using this weak supervision signal as a proxy for visual explanation generation. Unlike prior XAI approaches that depend on task-specific classification heads, *DeViL* is applicable to any vision backbone. It produces both natural language descriptions of individual feature map locations across different network layers and open-vocabulary saliency maps. We subsequently adapted *DeViL* for the medical domain by training on structured radiology reports and incorporating a radiology-specialised language model. Experiments on structured report generation, saliency evaluation, and saliency–text grounding across three types of CXR encoders showed competitive performance with large end-to-end report generation models and demonstrated that *DeViL*'s open-vocabulary saliency maps outperform those of a VLM-specific saliency method.

Overall, this thesis explored different explanation modalities (visual, concept-based, and textual) and weak explanation supervision paradigms (weight-input alignment, ICL, feature reconstruction and distribution matching, and image-caption alignment). Thus, we believe that, with the works here presented, we have achieved the goals we had set out from the start, i.e., propose alternatives to *post-hoc* visual explanations without requiring full supervision of the explanation generation processes and, therefore, without increasing the already costly annotation burden of medical image classification models. The scientific outputs resulting from this thesis have been disseminated in three international journal and five international conference publications.

Naturally, every work has limitations and we identify the lack of clinical validation as this work's main one. Although we tried to mitigate this by validating our proposals on medical datasets with clinician-annotated data using several explanation quality assessment metrics, validation with radiologists is needed before any of the proposed models are ready to be deployed into practice.

Evaluating explanations is, in itself, more difficult than evaluating, e.g., classification performance, mainly due to the subjectivity of the problem (two people might explain the same thing in completely different, albeit correct, ways). Furthermore, methods should explain the underlying models' decisions and not aim to mimic human annotations. Although these issues are not easily solvable and, thus, remain open challenges of the field, the evaluation of XAI methods is made worse by something we have come across throughout the development of this thesis and that we consider to be one of the major problems of the XAI field: the lack of standardisation (and this applies to any modality of explanation). First of all, different authors report different metrics, making it difficult to compare approaches. To mitigate this problem we report several metrics and try to evaluate different properties of the generated explanations. Secondly, some authors report metrics for all instances of the test set, while others only consider correctly predicted cases. Working with multilabel data, where labels are not mutually exclusive, adds to the issue, because there is no consensus on how to explain this type of data (e.g., provide one single explanation per instance, taking into account all predicted labels, or provide one explanation per label per instance). Finally, explanation methods usually focus on explaining instances in which a pathology has been diagnosed, ignoring instances of healthy cases. This happens in part because the few existing datasets usually do not have GT explanations for the latter and because in many applications it is not clear how to explain a healthy instance. In fact, that is something that we want to explore in the future, especially regarding the generation of NLEs with *WeNLEX*.

Despite the aforementioned challenges (mainly lack of clinical validation), we believe the contributions of this thesis highlight the importance of and pave the way for more versatile clinical decision support systems that leverage the complementarity of different explanation modalities. Thus, an important and promising future research direction is the design of end-to-end frameworks capable of dynamically adapting the type and depth of their explanations to the needs of each user (e.g., offering a radiologist a spatially precise saliency map alongside a structured clinical report, while presenting a patient with an accessible natural language summary). However, such versatility cannot be achieved at the expense of an ever-growing annotation burden; future methods should continue to strive to learn to generate explanations with minimal supervision, not only due to the scarcity and cost of expert clinical annotations, but to ensure explanations reflect the true decision processes of the underlying models and not the human-provided annotations. In fact, the proposal of *WeNLEX* constitutes a solid step in this direction and we plan on integrating visual explanations into the framework to increase its adaptability.

References

[1] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLOS Medicine*, vol. 15, pp. 1–4, 2018.

[2] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, "Understanding the decisions of CNNs: An in-model approach," *Pattern Recognition Letters*, vol. 133, pp. 373–380, 2020.

[3] M. Böhle, M. Fritz, and B. Schiele, "B-cos Networks: Alignment is All We Need for Interpretability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 319–10 328.

[4] M. Böhle, N. Singh, M. Fritz, and B. Schiele, "B-Cos Alignment for Inherently Interpretable CNNs and Vision Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 46, no. 6, pp. 4504–4518, 2024.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[6] T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, "A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?" *IEEE Access*, vol. 10, pp. 98 909–98 935, 2022.

[7] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9525–9536.

[9] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom, "Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods," in *Proceedings of the NeurIPS Workshop on Safety and Robustness in Decision Making*, 2019.

[10] F. Sammani and N. Deligiannis, "Zero-Shot Natural Language Explanations," in *International Conference on Learning Representations (ICLR)*, 2025.

[11] Karthik, Maggie, and S. Dane, *APTOS 2019 Blindness Detection*, https://kaggle.com/competitions/aptos2019-blindness-detection, Accessed: 2025-04-07, 2019.

[12] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific Data*, vol. 8, no. 1, pp. 1–8, 2021.

[13] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.

[14] I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, "On the Suitability of B-cos Networks for the Medical Domain," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[15] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification," in *Proceedings of Nature inspired Smart Information Systems (NiSIS)*, 2005, pp. 1–9.

[16] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180 161, 2018.

[17] N. Codella, V. Rotemberg, P. Tschandl, *et al.*, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1902.03368*, 2019.

[18] H. Q. Nguyen, K. Lam, L. T. Le, *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, p. 429, 2022.

[19] M. Alesina, C. A. Barbano, C. Berzovini, *et al.*, *CORDA Dataset*, version 2.0, Zenodo, 2023. DOI: 10.5281/zenodo.7821611.

[20] C. Patrício, I. Rio-Torto, J. S. Cardoso, L. F. Teixeira, and J. C. Neves, "CBVLM: Training-free explainable concept-based Large Vision Language Models for medical image classification," *Computers in Biology and Medicine*, vol. 198, p. 111 145, 2025.

[21] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," *IEEE Journal of Biomedical and Health Informatics (JBHI)*, vol. 23, no. 2, pp. 538–546, 2019.

[22] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019.

[23] R. Daneshjou, M. Yuksekgonul, Z. R. Cai, R. Novoa, and J. Y. Zou, "SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 18 157–18 167.

[24] M. Kayser, C. Emde, O.-M. Camburu, G. Parsons, B. Papiez, and T. Lukasiewicz, "Explaining Chest X-Ray Pathologies in Natural Language," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science (LNCS), 2022, pp. 701–713.

[25] I. Rio-Torto, J. S. Cardoso, and L. F. Teixeira, "From Captions to Explanations: A Multimodal Transformer-based Architecture for Natural Language Explanation Generation," in *Pattern Recognition and Image Analysis*, 2022, pp. 54–65.

[26] I. Rio-Torto, J. S. Cardoso, and L. F. Teixeira, "Parameter-Efficient Generation of Natural Language Explanations for Chest X-ray Classification," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 250, 2024, pp. 1267–1281.

[27] I. Rio-Torto, J. S. Cardoso, and L. F. Teixeira, "WeNLEX: Weakly Supervised Natural Language Explanations for Multilabel Chest X-ray Classification," *under revision at IEEE Transactions on Medical Imaging (TMI)*, 2025.

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

[29] M. Dani, I. Rio-Torto, S. Alaniz, and Z. Akata, "DeViL: Decoding Vision features into Language," in *Proceedings of the DAGM German Conference on Pattern Recognition (GCPR)*, 2023, pp. 363–377.

[30] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2014, pp. 740–755.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 6, pp. 1452–1464, 2018.

[33] E. Hernandez, S. Schwettmann, D. Bau, T. Bagashvili, A. Torralba, and J. Andreas, "Natural Language Descriptions of Deep Features," in *International Conference on Learning Representations (ICLR)*, 2022.

[34] J.-B. Delbrouck, J. Xu, J. Moll, *et al.*, "Automated Structured Radiology Report Generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 26 813–26 829.

[35] I. Rio-Torto, B. Coelho, J. S. Cardoso, and L. F. Teixeira, "Understanding Chest X-ray Vision Representations with DeViL," *accepted at the IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2025.

[36] A. Saporta, X. Gui, A. Agrawal, *et al.*, "Benchmarking saliency methods for chest X-ray interpretation," *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022.

[37] B. Boecking, N. Usuyama, S. Bannur, *et al.*, "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2022, pp. 1–21.

[38] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," in *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience*. Springer Netherlands, 1987, pp. 115–141.

[39] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[40] S. Minut and S. Mahadevan, "A Reinforcement Learning Model of Selective Visual Attention," in *Proceedings of the International Conference on Autonomous Agents*, 2001, pp. 457–464.

[41] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[43] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[44] S. Jetley, N. A. Lord, N. Lee, and P. Torr, "Learn to Pay Attention," in *International Conference on Learning Representations (ICLR)*, 2018.

[45] I. Makarov, M. Bakhanova, S. Nikolenko, and O. Gerasimova, "Self-supervised recurrent depth estimation with attention mechanisms," *PeerJ Computer Science*, vol. 8, e865, 2022.

[46] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6000–6010.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[49] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An Attentive Survey of Attention Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 5, 2021.

[50] K. Xu, J. Ba, R. Kiros, *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), 2015, pp. 2048–2057.

[51] M.-H. Guo, T.-X. Xu, J.-J. Liu, *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, 2022.

[52] L. Chen, H. Zhang, J. Xiao, *et al.*, "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5659–5667.

[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2018, pp. 3–19.

[55] B. Wu, C. Xu, X. Dai, *et al.*, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," *arXiv:2006.03677*, 2020.

[56] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultrasound Images," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022, pp. 480–483.

[57] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 80, 2018, pp. 2127–2136.

[58] Z. Han, B. Wei, Y. Hong, *et al.*, "Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning," *IEEE Transactions on Medical Imaging (TMI)*, vol. 39, no. 8, pp. 2584–2594, 2020.

[59] S. Perera, S. Adhikari, and A. Yilmaz, "POCFormer: A Lightweight Transformer Architecture for Detection of COVID-19 Using Point of Care Ultrasound," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 195–199.

[60] J. Jiang and S. Lin, "COVID-19 Detection in Chest X-ray Images Using Swin-Transformer and Transformer in Transformer," *arXiv:2110.08427*, 2021.

[61] C. Liu and Q. Yin, "Automatic Diagnosis of COVID-19 Using a tailored Transformer-Like Network," in *Journal of Physics: Conference Series*, vol. 2010, 2021, p. 012 175.

[62] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated Split Task-Agnostic Vision Transformer for COVID-19 CXR Diagnosis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 24 617–24 630.

[63] C.-C. Hsu, G.-L. Chen, and M.-H. Wu, "Visual Transformer with Statistical Test for COVID-19 Classification," *arXiv:2107.05334*, 2021.

[64] L. Zhang and Y. Wen, "A transformer-based framework for automatic COVID19 diagnosis in chest CTs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 513–518.

[65] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. Prathosh, "xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–10, 2022.

[66] X. Gao, M. H.-M. Khan, R. Hui, *et al.*, "COVID-VIT: Classification of Covid-19 from 3D CT chest images based on vision transformer model," in *Proceedings of the International Conference on Next Generation Computing Applications (NextComp)*, 2022, pp. 1–4.

[67] D. Shome, T. Kar, S. N. Mohanty, *et al.*, "COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11 086, 2021.

[68] A. A. E. Ambita, E. N. V. Boquio, and P. C. Naval, "COViT-GAN: Vision Transformer for COVID-19 Detection in CT Scan Images with Self-Attention GAN for Data Augmentation," in *International Conference on Artificial Neural Networks*, 2021, pp. 587–598.

[69] S. Park, G. Kim, Y. Oh, *et al.*, "Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification," *Medical Image Analysis*, vol. 75, p. 102 299, 2022.

[70] M. Lu, Y. Pan, D. Nie, *et al.*, "SMILE: Sparse-Attention based Multiple Instance Contrastive Learning for Glioma Sub-Type Classification Using Pathological Images," in *MICCAI Workshop on Computational Pathology*, ser. Proceedings of Machine Learning Research (PMLR), 2021, pp. 159–169.

[71] H. Chen, C. Li, G. Wang, *et al.*, "GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection," *Pattern Recognition*, vol. 130, p. 108 827, 2022.

[72] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for Diagnosis of Acute Lymphoblastic Leukemia Based on ViT-CNN Ensemble Model," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[73] Y. Zheng, R. H. Gindra, E. J. Green, *et al.*, "A Graph-Transformer for Whole Slide Image Classification," *IEEE Transactions on Medical Imaging (TMI)*, vol. 41, no. 11, pp. 3003–3015, 2022.

[74] J. D. Bodapati, N. S. Shaik, and V. Naralasetti, "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.

[75] S. Yu, K. Ma, Q. Bi, *et al.*, "MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science (LNCS), 2021, pp. 45–54.

[76] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, "Lesion-Aware Transformers for Diabetic Retinopathy Grading," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 938–10 947.

[77] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision Transformer-based recognition of diabetic retinopathy grade," *Medical Physics*, vol. 48, no. 12, pp. 7850–7863, 2021.

[78] N. AlDahoul, H. Abdul Karim, M. Joshua Toledo Tan, M. A. Momo, and J. Ledesma Fermin, "Encoding Retina Image to Words using Ensemble of Vision Transformers for Diabetic Retinopathy Grading," *F1000Research*, vol. 10, p. 948, 2021.

[79] H. Yang, J. Chen, and M. Xu, "Fundus Disease Image Classification based on Improved Transformer," in *International Conference on Neuromorphic Computing (ICNC)*, 2021, pp. 207–214.

[80] M. Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognition*, vol. 122, p. 108 309, 2022.

[81] J. Moranguinho, T. Pereira, B. Ramos, J. Morgado, J. L. Costa, and H. P. Oliveira, "Attention Based Deep Multiple Instance Learning Approach for Lung Cancer Prediction using Histopathological Images," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021, pp. 2852–2855.

[82] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.

[83] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers Advance Multi-Modal Medical Image Classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[84] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft Attention Improves Skin Cancer Classification Performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*, 2021, pp. 13–23.

[85] Y. Barhoumi and G. Rasool, "Scopeformer: n-CNN-ViT Hybrid Model for Intracranial Hemorrhage Classification," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), 2021.

[86] S. Liang and Y. Gu, "Computer-Aided Diagnosis of Alzheimer's Disease through Weak Supervision Deep Learning Framework with Attention Mechanism," *Sensors*, vol. 21, no. 1, p. 220, 2021.

[87] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[88] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[89] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[90] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), 2021, pp. 10 347–10 357.

[91] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 70, 2017, pp. 3145–3153.

[92] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[93] N. Kokhlikyan, V. Miglani, M. Martin, *et al.*, "Captum: A unified and generic model interpretability library for PyTorch," *arXiv:2009.07896*, 2020.

[94] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 782–791.

[95] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the Relationship between Self-Attention and Convolutional Layers," in *International Conference on Learning Representations (ICLR)*, 2020.

[96] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[97] Z. C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[98] P.-J. Kindermans, S. Hooker, J. Adebayo, *et al.*, "The (Un)reliability of Saliency Methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer International Publishing, 2019, pp. 267–280.

[99] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[100] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[101] A. Hedström, L. Weber, D. Bareeva, *et al.*, "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond," *Journal of Machine Learning Research (JMLR)*, vol. 24, no. 1, 2023.

[102] Z. Liu, R. Xiong, and T. Jiang, "CI-Net: Clinical-Inspired Network for Automated Skin Lesion Recognition," *IEEE Transactions on Medical Imaging (TMI)*, vol. 42, no. 3, pp. 619–632, 2023.

[103] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *British Machine Vision Conference (BMVC)*, 2018, p. 151.

[104] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[105] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and Aggregating Feature-based Model Explanations," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[106] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (In)fidelity and Sensitivity of Explanations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[107] C. Agarwal, N. Johnson, M. Pawelczyk, *et al.*, "Rethinking Stability for Attribution-based Explanations," in *International Conference on Learning Representations Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data (ICLRW)*, 2022.

[108] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise Explanations of Neural Networks using Adversarial Training," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 119, 2020, pp. 1383–1391.

[109] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Giménez-Ábalos, "Focus! Rating XAI Methods and Finding Biases," in *FUZZ-IEEE*, 2022, pp. 1–8.

[110] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards Best Practice in Explaining Neural Network Decisions with LRP," in *IJCNN*, 2020, pp. 1–7.

[111] C. Matsoukas, J. F. Haslum, M. Sorkhei, M. Söderberg, and K. Smith, "What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9215–9224.

[112] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 70, 2017, pp. 3319–3328.

[113] M. Moor, Q. Huang, S. Wu, *et al.*, "Med-Flamingo: a Multimodal Medical Few-shot Learner," in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research (PMLR), vol. 225, 2023, pp. 353–367.

[114] C. Li, C. Wong, S. Zhang, *et al.*, "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[115] Z. Chen, M. Varma, J.-B. Delbrouck, *et al.*, "CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[116] J. Zhou, X. He, L. Sun, *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024.

[117] T. Han, L. C. Adams, S. Nebelung, J. N. Kather, K. K. Bressem, and D. Truhn, "Multimodal Large Language Models are Generalist Medical Image Interpreters," *medRxiv*, pp. 2023–12, 2023.

[118] M.-H. Van, P. Verma, and X. Wu, "On Large Visual Language Models for Medical Imaging Analysis: An Empirical Study," in *Proceedings of the IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2024, pp. 172–176.

[119] D. Ferber, G. Wölflein, I. C. Wiest, *et al.*, "In-context learning enables multimodal large language models to classify cancer pathology images," *Nature Communications*, vol. 15, no. 1, p. 10 104, 2024.

[120] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 34 892–34 916.

[121] OpenAI, "GPT-4 Technical Report," *arXiv:2303.08774*, 2023.

[122] C. Royer, B. Menze, and A. Sekuboyina, "MultiMedEval: A Benchmark and a Toolkit for Evaluating Medical Vision-Language Models," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 250, 2024, pp. 1310–1327.

[123] J. Liu, W. Wang, S. Yihang, *et al.*, "Asclepius: A Spectrum Evaluation Benchmark for Medical Multi-Modal Large Language Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 24 181–24 201.

[124] P. Xia, Z. Chen, J. Tian, *et al.*, "CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 140 334–140 365.

[125] P. W. Koh, T. Nguyen, Y. S. Tang, *et al.*, "Concept Bottleneck Models," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), 2020, pp. 5338–5348.

[126] C. Patrício, J. C. Neves, and L. F. Teixeira, "Coherent Concept-Based Explanations in Medical Image and Its Application to Skin Lesion Diagnosis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3798–3807.

[127] Y. Bie, L. Luo, and H. Chen, "MICA: Towards Explainable Skin Lesion Diagnosis via Multi-Level Image-Concept Alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 837–845.

[128] C. Patrício, L. F. Teixeira, and J. C. Neves, "Towards Concept-based Interpretability of Skin Lesion Diagnosis using Vision-Language Models," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[129] S. Ghosh, K. Yu, F. Arabshahi, and K. Batmanghelich, "Bridging the Gap: From Post Hoc Explanations to Inherently Interpretable Models for Medical Imaging," in *Proceedings of the ICML Workshop on Interpretable Machine Learning in Healthcare (ICMLW)*, 2023.

[130] OpenAI, *ChatGPT: Optimizing Language Models for Dialogue*, https://chat.openai.com/, 2022.

[131] T. Brown, B. Mann, N. Ryder, *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[132] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 23 716–23 736.

[133] G. Zhou, Z. Han, S. Chen, *et al.*, "Adapting Large Multimodal Models to Distribution Shifts: The Role of In-Context Learning," *arXiv:2405.12217*, 2024.

[134] A. Awadalla, I. Gao, J. Gardner, *et al.*, "OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models," *arXiv:2308.01390*, 2023.

[135] Z. Yang, Z. Gan, J. Wang, *et al.*, "An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 3081–3089.

[136] S. Chen, Z. Han, B. He, *et al.*, "Understanding and Improving In-Context Learning on Vision-language Models," in *International Conference on Learning Representations Workshop on Mathematical and Empirical Understanding of Foundation Models (ICLRW)*, 2024.

[137] X. Wang, B. Ma, C. Hu, *et al.*, ""My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models," in *Findings of the Association for Computational Linguistics (ACL)*, 2024, pp. 7407–7416.

[138] M. Groh, C. Harris, L. Soenksen, *et al.*, "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1820–1828.

[139] C. A. Barbano, E. Tartaglione, C. Berzovini, M. Calandri, and M. Grangetto, "A Two-Step Radiologist-Like Approach for Covid-19 Computer-Aided Diagnosis from Chest X-Ray Images," in *Image Analysis and Processing – ICIAP 2022*, 2022, pp. 173–184.

[140] J. Irvin, P. Rajpurkar, M. Ko, *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[141] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon, "Building and better understanding vision-language models: insights and future directions," in *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

[142] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "VILA: On Pre-training for Visual Language Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 689–26 699.

[143] B. Li, Y. Zhang, D. Guo, *et al.*, "LLaVA-OneVision: Easy Visual Task Transfer," *Transactions on Machine Learning Research (TMLR)*, 2025.

[144] J. Bai, S. Bai, S. Yang, *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," *arXiv:2308.12966*, 2023.

[145] Y. Yao, T. Yu, A. Zhang, *et al.*, "MiniCPM-V: A GPT-4V Level MLLM on Your Phone," *arXiv:2408.01800*, 2024.

[146] Z. Chen, W. Wang, Y. Cao, *et al.*, "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling," *arXiv:2412.05271*, 2025.

[147] J. Ye, H. Xu, H. Liu, *et al.*, "mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models," in *International Conference on Learning Representations (ICLR)*, vol. 2025, 2025, pp. 98 891–98 913.

[148] C. Wen, M. Ye, H. Li, T. Chen, and X. Xiao, "Concept-Based Lesion Aware Transformer for Interpretable Retinal Disease Diagnosis," *IEEE Transactions on Medical Imaging (TMI)*, vol. 44, no. 1, pp. 57–68, 2025.

[149] J. Silva-Rodríguez, H. Chakor, R. Kobbi, J. Dolz, and I. Ben Ayed, "A Foundation Language-Image Model of the Retina (FLAIR): encoding expert knowledge in text supervision," *Medical Image Analysis*, vol. 99, p. 103 357, 2025.

[150] N. C. Codella, Y. Jin, S. Jain, *et al.*, "MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging," *arXiv:2410.06542*, 2024.

[151] J. Hou, J. Xu, and H. Chen, "Concept-Attention Whitening for Interpretable Skin Lesion Diagnosis," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science (LNCS), 2024, pp. 113–123.

[152] C. A. Barbano, R. Renzulli, M. Grosso, D. Basile, M. Busso, and M. Grangetto, "AI-Assisted Diagnosis for Covid-19 CXR Screening: from Data Collection to Clinical Validation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–4.

[153] S. Madarapu, S. Ari, and K. Mahapatra, "A multi-resolution convolutional attention network for efficient diabetic retinopathy classification," *Computers and Electrical Engineering*, vol. 117, p. 109 243, 2024.

[154] S. Zhang, Y. Xu, N. Usuyama, *et al.*, "A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs," *NEJM AI*, vol. 2, no. 1, AIoa2400640, 2025.

[155] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved Training Techniques for CLIP at Scale," *arXiv:2303.15389*, 2023.

[156] K. Chauhan, R. Tiwari, J. Freyberg, P. Shenoy, and K. Dvijotham, "Interactive Concept Bottleneck Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5948–5955.

[157] T. Li, G. Zhang, Q. D. Do, X. Yue, and W. Chen, "Long-context LLMs Struggle with Long In-context Learning," *Transactions on Machine Learning Research (TMLR)*, 2025.

[158] H. Zhang, W. Shao, H. Liu, *et al.*, "B-AVIBench: Toward Evaluating the Robustness of Large Vision-Language Model on Black-Box Adversarial Visual-Instructions," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 1434–1446, 2025.

[159] M. Phung, V. Muralidharan, V. Rotemberg, *et al.*, "Best Practices for Clinical Skin Image Acquisition in Translational Artificial Intelligence Research," *Journal of Investigative Dermatology*, vol. 143, no. 7, pp. 1127–1132, 2023.

[160] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[161] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Producing radiologist-quality reports for interpretable artificial intelligence," *arXiv:1806.00340*, 2018.

[162] C. Agarwal, S. H. Tanneru, and H. Lakkaraju, "Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models," *arXiv:2402.04614*, 2024.

[163] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the Visualization of What a Deep Neural Network Has Learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[164] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating Visual Explanations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2016, pp. 3–19.

[165] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 632–642.

[166] D. H. Park, L. A. Hendricks, Z. Akata, *et al.*, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8779–8788.

[167] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual Entailment: A Novel Task for Fine-grained Image Understanding," *arXiv:1901.06706*, 2019.

[168] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-SNLI: Natural Language Inference with Natural Language Explanations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9560–9572.

[169] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 4932–4942.

[170] S. Kumar and P. Talukdar, "NILE: Natural Language Inference with Faithful Natural Language," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 8730–8742.

[171] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, and K. Malkan, "WT5?! Training Text-to-Text Models to Explain their Predictions," *arXiv:2004.14546*, 2020.

[172] J. Wu and R. Mooney, "Faithful Multimodal Explanation for Visual Question Answering," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 103–112.

[173] V. Do, O.-M. Camburu, Z. Akata, and T. Lukasiewicz, "e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations," *arXiv:2004.03744*, 2020.

[174] A. Marasović, C. Bhagavatula, J. s. Park, R. Le Bras, N. A. Smith, and Y. Choi, "Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graph," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2810–2829.

[175] M. Kayser, O.-M. Camburu, L. Salewski, *et al.*, "e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1224–1234.

[176] B. P. Majumder, O. Camburu, T. Lukasiewicz, and J. Mcauley, "Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 162, 2022, pp. 14 786–14 801.

[177] B. Plüster, J. Ambsdorf, L. Braach, J. H. Lee, and S. Wermter, "Harnessing the Power of Multi-Task Pretraining for Ground-Truth Level Natural Language Explanations," *arXiv:22 12.04231*, 2023.

[178] F. Sammani, T. Mukherjee, and N. Deligiannis, "NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8312–8322.

[179] E. Zablocki, V. Gerard, A. Cardiel, E. Gaussier, M. Cord, and E. Valle, "GIFT: A Framework Towards Global Interpretable Faithful Textual Explanations of Vision Classifiers," *Transactions on Machine Learning Research (TMLR)*, 2026.

[180] A. Wojciechowski, M. Lango, and O. Dusek, "Faithful and Plausible Natural Language Explanations for Image Classification: A Pipeline Approach," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 2340–2351.

[181] F. Sammani and N. Deligiannis, "Uni-NLX: Unifying Textual Explanations for Vision and Vision-Language Tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 4636–4641.

[182] A. Hamza, A. Abdullah, Y. H. Ahn, S. Lee, and S. T. Kim, "LLaVA Needs More Knowledge: Retrieval Augmented Natural Language Generation with Knowledge Graph for Explaining Thoracic Pathologies," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, pp. 3311–3319, 2025.

[183] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science (LNCS), 2021, pp. 293–303.

[184] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language Models Are Unsupervised Multitask Learners," *OpenAI Technical Report*, 2019.

[185] Y. Yu, C.-H. H. Yang, J. Kolehmainen, *et al.*, "Low-Rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.

[186] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 296–26 306.

[187] S. Jain, A. Agrawal, A. Saporta, *et al.*, "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports," in *NeurIPS Track on Datasets and Benchmarks*, vol. 1, 2021.

[188] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 3876–3887.

[189] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3557–3567.

[190] C. Wang, S. Liang, Y. Jin, Y. Wang, X. Zhu, and Y. Zhang, "SemEval-2020 Task 4: Commonsense Validation and Explanation," in *14th Workshop on Semantic Evaluation*, 2020, pp. 307–321.

[191] S. Aggarwal, D. Mandowara, V. Agrawal, D. Khandelwal, P. Singla, and D. Garg, "Explanations for CommonsenseQA: New Dataset and Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 3050–3065.

[192] S. Wiegreffe and A. Marasović, "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing," in *NeurIPS Track on Datasets and Benchmarks*, vol. 1, 2021.

[193] P. Chambon, J.-B. Delbrouck, T. Sounack, *et al.*, "CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats," *arXiv:2405.19538*, 2024.

[194] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101 797, 2020.

[195] M. D. L. I. Vayá, J. M. Saborit, J. A. Montell, *et al.*, "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients," *arXiv:2006.01174*, 2020.

[196] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, and M. Lungren, "Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1500–1519.

[197] D. Donahue and A. Rumshisky, "Adversarial Text Generation Without Reinforcement Learning," *arXiv:1810.06640*, 2018.

[198] I. Montero, N. Pappas, and N. A. Smith, "Sentence Bottleneck Autoencoders from Transformer Language Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 1822–1831.

[199] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 506–516.

[200] N. Houlsby, A. Giurgiu, S. Jastrzebski, *et al.*, "Parameter-Efficient Transfer Learning for NLP," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 97, 2019, pp. 2790–2799.

[201] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 3045–3059.

[202] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) and the International Joint Conference on Natural Language Processing*, 2021, pp. 4582–4597.

[203] X. Liu, K. Ji, Y. Fu, *et al.*, "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 61–68.

[204] Z. Chen, Y. Duan, W. Wang, *et al.*, "Vision Transformer Adapter for Dense Predictions," in *International Conference on Learning Representations (ICLR)*, 2023.

[205] R. Zhang, J. Han, C. Liu, *et al.*, "LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention," in *International Conference on Learning Representations (ICLR)*, 2024.

[206] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, and R. Ji, "Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 29 615–29 627.

[207] P. Gao, S. Geng, R. Zhang, *et al.*, "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[208] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, and T. Hospedales, "Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 250, 2024, pp. 406–425.

[209] J. Solano, M. Sanni, O.-M. Camburu, and P. Minervini, "SparseFit: Few-shot Prompting with Sparse Fine-tuning for Jointly Generating Predictions and Natural Language Explanations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 2053–2077.

[210] H. Touvron, T. Lavril, G. Izacard, *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv:2302.13971*, 2023.

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

[212] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.

[213] U.S. National Library of Medicine, *Pubmed*, Accessed: 2025-05-19. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/.

[214] A. E. W. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160 035, 2016.

[215] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[216] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4198–4205.

[217] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), 2017, pp. 214–223.

[218] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5769–5779.

[219] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[220] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, 2021.

[221] J. Xu, Z. Chen, A. Johnston, *et al.*, "Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!"" In *Workshop on Biomedical NLP*, 2024, pp. 85–98.

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

[223] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608*, 2017.

[224] P. Hase, S. Zhang, H. Xie, and M. Bansal, "Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?" In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4351–4367.

[225] S. Wiegreffe, A. Marasović, and N. A. Smith, "Measuring Association Between Labels and Free-Text Rationales," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 10 266–10 284.

[226] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein, "Faithfulness Tests for Natural Language Explanations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023, pp. 283–294.

[227] N. Siegel, O.-M. Camburu, N. Heess, and M. Perez-Ortiz, "The Probabilities Also Matter: A More Faithful Metric for Faithfulness of Free-Text Explanations in Large Language Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 530–546.

[228] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[229] P. Atrey, M. P. Brundage, M. Wu, and S. Dutta, "Demystifying the Accuracy-Interpretability Trade-Off: A Case Study of Inferring Ratings from Reviews," *arXiv:2503.07914*, 2025.

[230] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2016, pp. 694–711.

[231] Z. Zhu, H. Jiang, J. Yang, *et al.*, "Situated Natural Language Explanations," *arXiv:2308.14115*, 2024.

[232] OpenAI, "GPT-4o System Card," *arXiv:2410.21276*, 2024.

[233] K. Zhao, C. Xiao, S. Yan, *et al.*, "X-ray Made Simple: Lay Radiology Report Generation and Robust Evaluation," *arXiv:2406.17911*, 2025.

[234] Z. Zhu, Z. Jin, J. Zhang, *et al.*, "Narrowing Information Bottleneck Theory for Multimodal Image-Text Representations Interpretability," in *International Conference on Learning Representations (ICLR)*, 2025.

[235] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), 2021, pp. 8748–8763.

[236] M. Tschannen, A. Gritsenko, X. Wang, *et al.*, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," *arXiv:2502.14786*, 2025.

[237] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going Deeper with Convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[238] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2014, pp. 818–833.

[239] P. Agrawal, J. Carreira, and J. Malik, "Learning to See by Moving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 37–45.

[240] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[241] W. Brendel and M. Bethge, "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet," in *International Conference on Learning Representations (ICLR)*, 2019.

[242] M. Böhle, M. Fritz, and B. Schiele, "Convolutional Dynamic Alignment Networks for Interpretable Classifications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 024–10 033.

[243] W. Liu, Z. Liu, Z. Yu, *et al.*, "Decoupled Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2771–2779.

[244] G. Zoumpourlis, A. Doumanoglou, N. Vretos, and P. Daras, "Non-linear Convolution Filters for CNN-Based Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 4771–4779.

[245] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4768–4777.

[246] B. Kim, M. Wattenberg, J. Gilmer, *et al.*, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 80, 2018, pp. 2668–2677.

[247] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[248] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[249] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3319–3327.

[250] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," *arXiv:2111.09734*, 2021.

[251] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 897–17 907.

[252] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 041–13 049, 2020.

[253] X. Hu, Z. Gan, J. Wang, *et al.*, "Scaling Up Vision-Language Pretraining for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 959–17 968.

[254] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research (JMLR)*, vol. 15, no. 1, pp. 1929–1958, 2014.

[255] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, pp. 376–380.

[256] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[257] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[258] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2016, pp. 382–398.

[259] S. Zhang, S. Roller, N. Goyal, *et al.*, "OPT: Open Pre-trained Transformer Language Models," *arXiv:2205.01068*, 2022.

[260] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[261] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 548–558.

[262] M. Ranjit, G. Ganapathy, S. Srivastav, T. Ganu, and S. Oruganti, "Rad-Phi2: Instruction Tuning Phi-2 for Radiology," *arXiv:2403.09725*, 2024.

[263] J. P. Cohen, J. D. Viviano, P. Bertin, *et al.*, "TorchXRayVision: A library of chest X-ray datasets and models," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), 2022, pp. 231–249.

[264] S. Feng, D. Azzollini, J. S. Kim, *et al.*, "Curation of the CANDID-PTX Dataset with Free-Text Reports," *Radiology: Artificial Intelligence*, vol. 3, no. 6, e210136, 2021.

[265] Z. Liu, H. Hu, Y. Lin, *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 009–12 019.

[266] S. Bannur, K. Bouzid, D. C. Castro, *et al.*, "MAIRA-2: Grounded Radiology Report Generation," *arXiv:2406.04449*, 2024.

[267] F. Pérez-García, H. Sharma, S. Bond-Taylor, *et al.*, "Exploring scalable medical image encoders beyond text supervision," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 119–130, 2025.

[268] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 46 595–46 623.

[269] J.-B. Delbrouck, P. Chambon, C. Bluethgen, E. Tsai, O. Almusa, and C. Langlotz, "Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 4348–4360.

[270] A. Osman, L. Arras, and W. Samek, "Towards Ground Truth Evaluation of Visual Explanations," *arXiv:2003.07258*, 2020.

[271] J. Wu, J. Wang, Z. Yang, *et al.*, "GRiT: A Generative Region-to-Text Transformer for Object Understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science (LNCS), 2024, pp. 207–224.

[272] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

Appendix A

Concept-based Explanations

Overview

In this appendix, we provide additional details that are not included in Chapter 3. Section A.1 presents additional ablation studies comparing example selection strategies, namely RICES [135] and Mixed Modality In-Context Example Selection (MMICES) [136]. Section A.2 offers further visualisations that illustrate the multidimensional interactions of concepts and model behavior in different LVLMs. Section A.3 includes details of the HuggingFace checkpoints for each adopted LVLM in our experiments. Section A.4 provides a list of the clinical concepts for each dataset. Section A.5 presents the detailed prompts used in our experiments for both concept detection and disease diagnosis tasks, across all datasets.

A.1 Comparison Between Example Selection Strategies - RICES and MMICES

In Chapter 3 all experiments are conducted using RICES [135] for the selection of the demonstration examples. This method selects the examples for ICL using the similarity between visual features. Chen et al [136] extend RICES to also take into account text features, in a method called MMICES, which the authors originally tested on vision-language datasets (e.g., image captioning, VQA). However, in our datasets such textual features do not exist. Thus, in the first stage of *CBVLM* (concept detection), MMICES cannot be applied. In the second stage (disease diagnosis), MMICES can be applied if we consider the concepts for each image as the textual features; the query image features are first compared to the train dataset image features, of which the K ($K = 200$) most similar are selected. Of these K example images, we select the n most similar to the query according to the cosine similarity between the concept vectors. The results of the comparison between the two approaches are presented in Figure A.1, which shows the consistent improvement of RICES over MMICES. This might be explained by the fact that MMICES was proposed for tasks where the textual features have greater inter-image variability (e.g., different image captions for different images), while in our use-cases different images (especially from the

same disease) might have the same set of concepts. Furthermore, we use a categorical representation of the set of concepts for each image, not a textual one. In the future it might be interesting to explore how to better adapt MMICES for this scenario or to use an image captioning model to provide textual descriptions of the images.

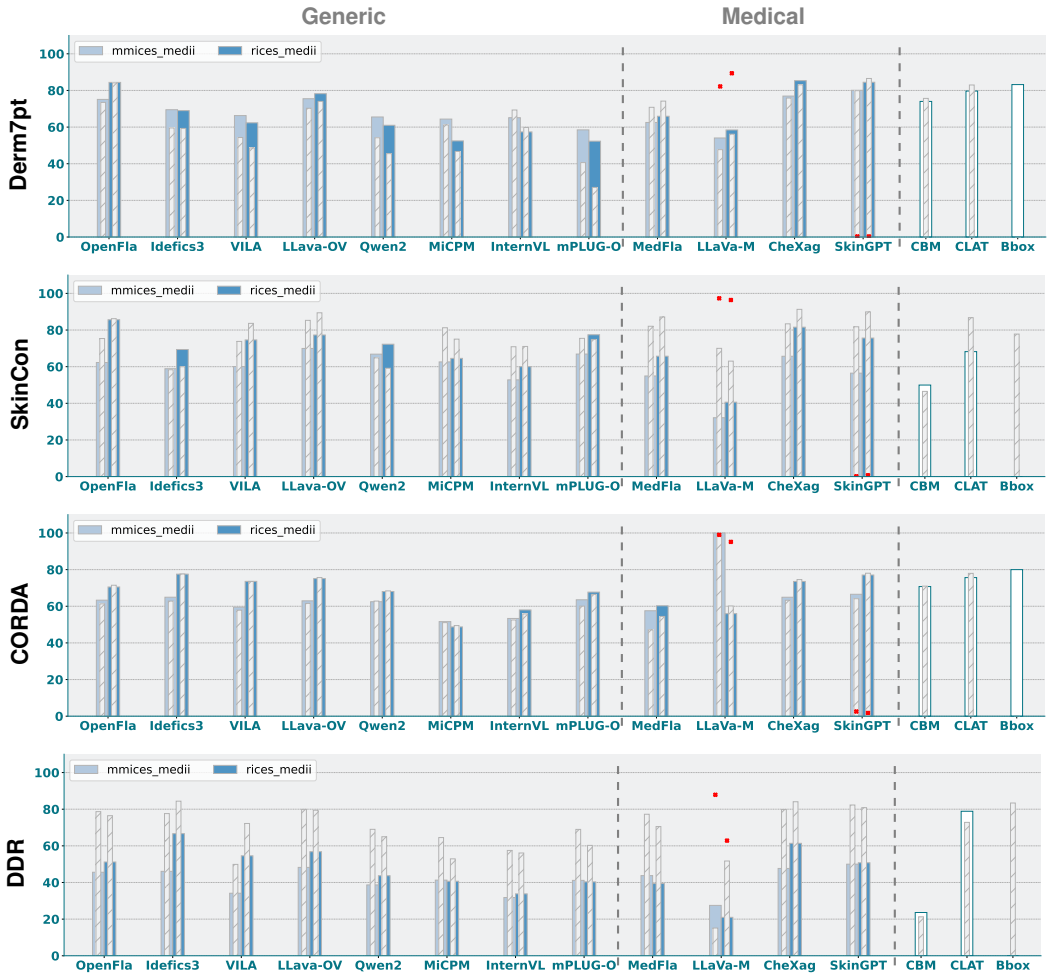


Figure A.1: Disease diagnosis results with 4-shots and two example selection strategies (RICES [135] and MMICES [136]). In both methodologies, MedImageInsight [150] is the vision encoder used to extract the features from the query and demonstration images.

A.2 Additional visualisations

Figures A.2 and A.3 present alternative visualisations to Figures 3.3 and 3.7, respectively, in the form of spider plots, illustrating model behavior across different datasets for the tasks of concept detection and disease diagnosis. It is interesting to note that LVLMs trained on a given domain are able to, without training and with just a few demonstration examples, perform competitively in completely different domains (*c.f.* CheXagent on skin datasets or SkinGPT-4 on CORDA).

Figure A.4 compares the per-concept performance of each evaluated LVLM in each of the four tested datasets, highlighting the LVLMs' ability to detect individual clinical concepts. As mentioned in the main text, and again can be verified here for each LVLM, underrepresented concepts are the ones for which the F1-scores are lower.

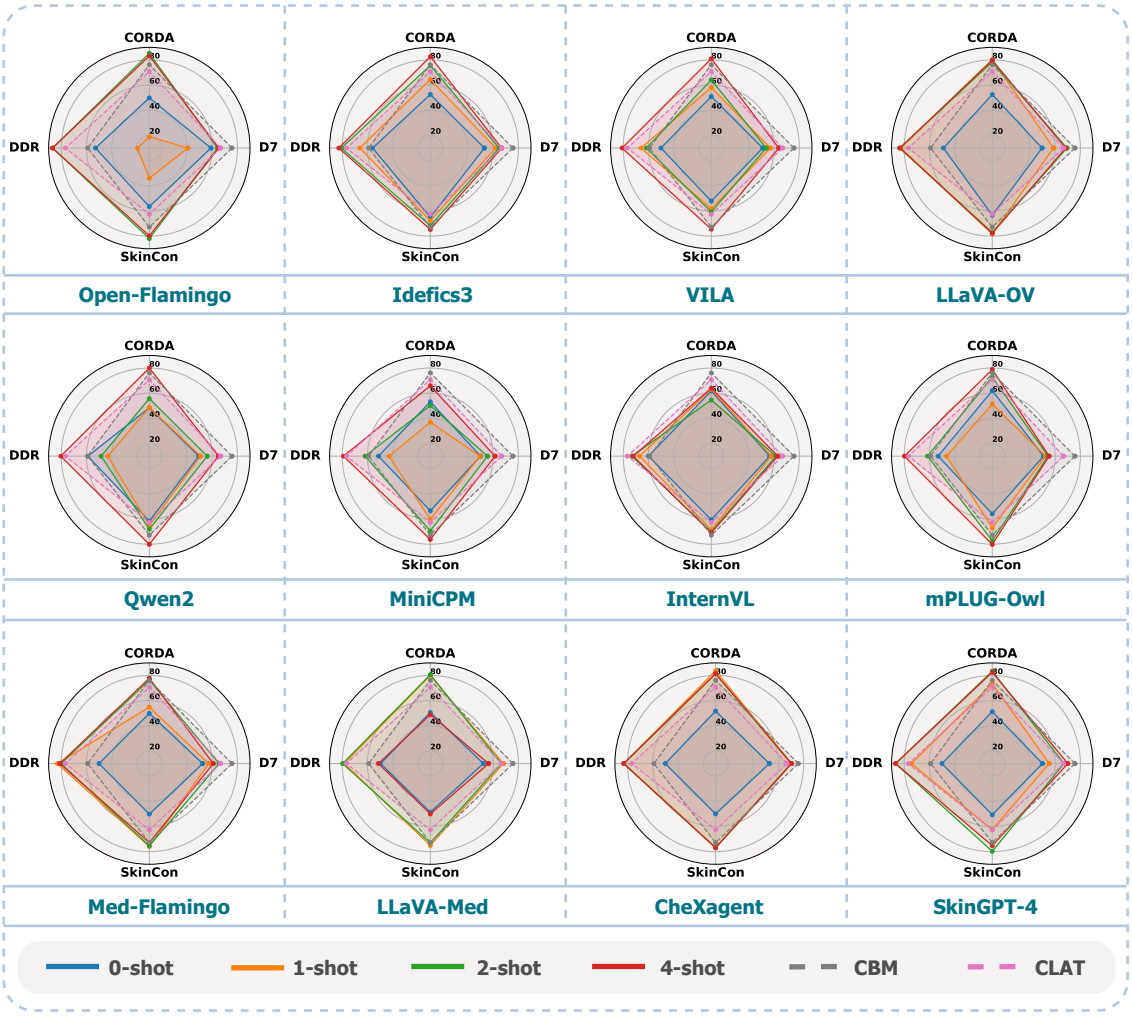


Figure A.2: Comparison of LVLM concept detection performance across datasets under different n-shot settings, reported in terms of BACC. Each n-shot setting is represented by a different colour. The gray and pink dashed lines indicate the performance of CBM and CLAT, respectively.

A.3 Implementation Details

The HuggingFace[1] checkpoints used for each LVLM are listed in Table A.1.

For SkinGPT-4 [116], we used the `skingpt4_llama2_13bchat_base_pre-train_stage2` checkpoint available at https://github.com/JoshuaChou2018/SkinGPT-4.

[1] https://huggingface.co

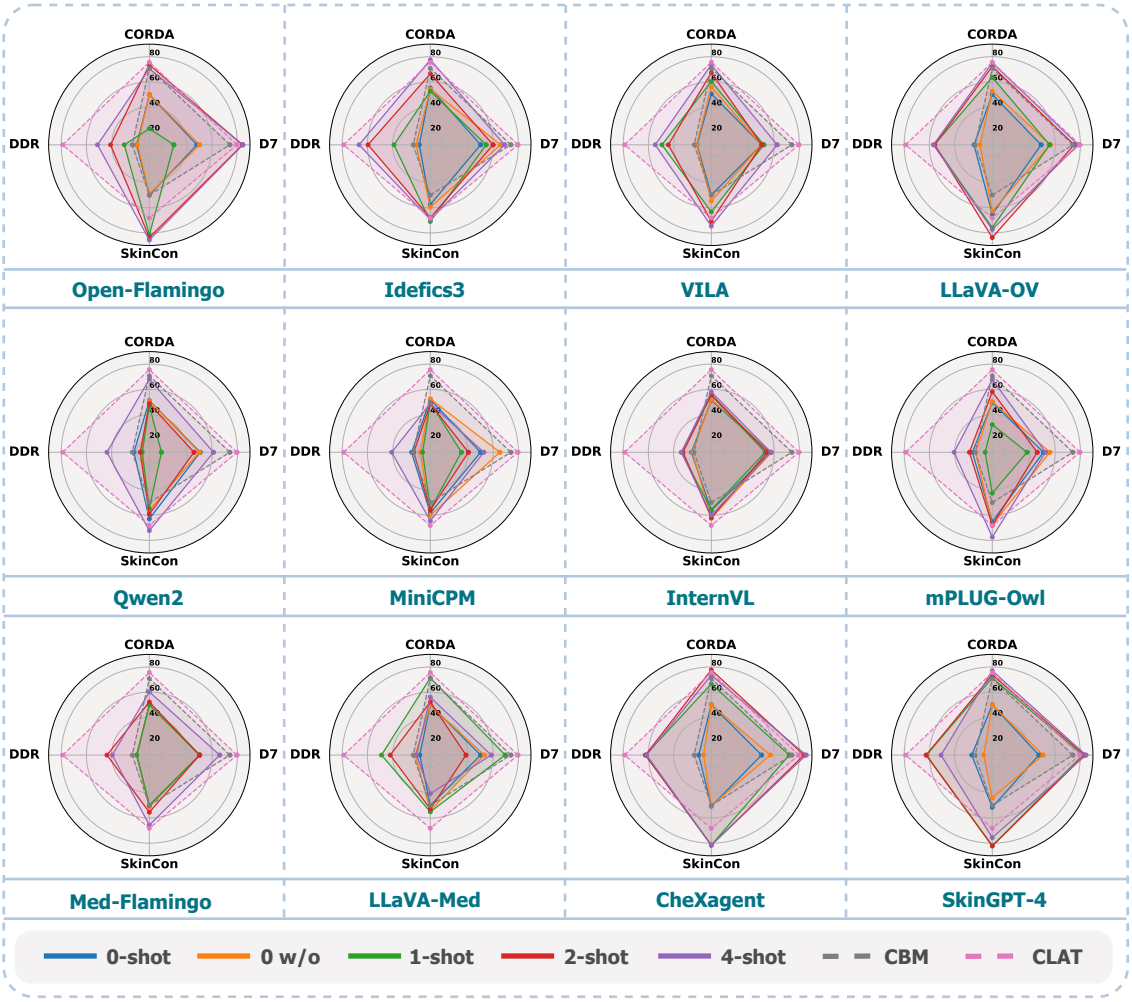


Figure A.3: Comparison of LVLM disease diagnosis performance across datasets under different n-shot settings, reported in terms of BACC. Each n-shot setting is represented by a different colour. The gray and pink dashed lines indicate the performance of CBM and CLAT, respectively.

For the automatic extraction of the LVLMs' answers we use the Mistral LLM, specifically the `mistralai/Mistral-7B-Instruct-v0.3` checkpoint at HuggingFace.

A.4 Clinical Concepts

Table A.2 lists all clinical concepts annotated in each dataset and used in our experiments. All concepts are binary (i.e., are either absent or present) except for "Pigment Network", "Streaks", and "Dots and Globules" of the Derm7pt dataset, for which the options are shown in the table.

A.5 Prompts

In the following section, we provide the prompts used in our experiments.

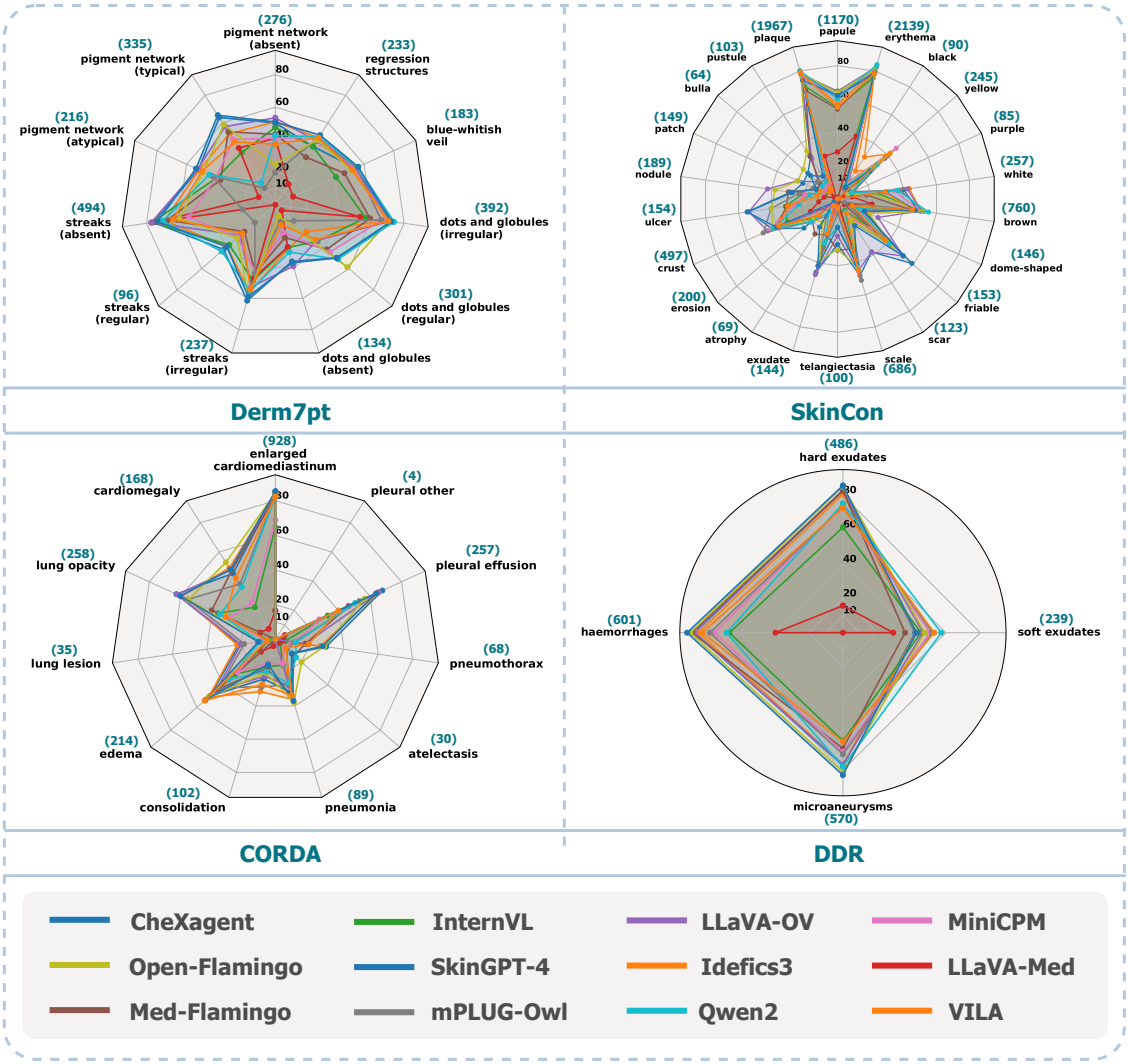


Figure A.4: Comparison of per-concept F1-scores across datasets for each evaluated LVLM, illustrating the relative performance patterns of the models in detecting individual clinical concepts. Values in parentheses indicate the number of samples in the dataset for the corresponding concept.

Concept Detection The template prompts used to query LVLMs in the task of concept detection are given in Figures A.5 to A.7 for each dataset.

Disease Diagnosis The template prompts used to query LVLMs in the task of disease diagnosis are given in Figures A.8 to A.11 for each dataset.

Table A.1: HuggingFace checkpoints of the LVLM used.

Model	Checkpoint
	Generic LVLMs
OpenFlamingo [134]	`openflamingo/OpenFlamingo-4B-vitl-rpj3b-langinstruct`
Idefics3 [141]	`HuggingFaceM4/Idefics3-8B-Llama3`
VILA [142]	`Efficient-Large-Model/Llama-3-VILA1.5-8B`
LLaVA-OneVision [143]	`lmms-lab/llava-onevision-qwen2-7b-ov`
Qwen2-VL [144]	`Qwen/Qwen2-VL-7B-Instruct`
MiniCPM-V [145]	`openbmb/MiniCPM-V-2_6`
InternVL 2.5 [146]	`OpenGVLab/InternVL2_5-8B`
mPLUG-Owl3 [147]	`mPLUG/mPLUG-Owl3-7B-240728`
	Medical LVLMs
Med-Flamingo [113]	`med-flamingo/med-flamingo`
LLaVA-Med [114]	`microsoft/llava-med-v1.5-mistral-7b`
CheXagent [115]	`StanfordAIMI/CheXagent-8b`

Instruction
The pigment network consists of intersecting brown lines forming a grid-like reticular pattern. It can be absent, typical, or atypical. A typical pigment network appears as a regular grid-like pattern on dermoscopy, consisting of thin lines that form an even mesh. The spaces (holes) between these lines are relatively uniform in size and shape. In an atypical pigment network the lines forming the network are uneven in thickness, and the holes or spaces between the lines vary in size and shape.

Demonstrations
Consider the following examples:
In the <image>, the pigment network is:
A) Absent
B) Typical
C) Atypical
Choose one option. Do not provide additional information.
Answer: A) Absent
(...)

Query
In the <image>, the pigment network is:
A) Typical
B) Atypical
C) Absent
Choose one option. Do not provide additional information.
Answer:

Figure A.5: Example concept detection prompt for the Derm7pt dataset with a multiple-choice concept ("pigment network"). The same template applies to the "streaks" and "dot and globules" concepts, with the corresponding instruction. The instructions for all Derm7pt concepts can be found in Figure A.8. Only 1 demonstration is shown.

Table A.2: List of annotated clinical concepts of each dataset.

Dataset	Clinical Concepts
Derm7pt	Pigment Network - {Absent, Typical, Atypical} Streaks - {Absent, Regular, Irregular} Dots and Globules - {Absent, Regular, Irregular} Blue-Whitish Veil Regression Structures
SkinCon	Papule, Erythema Plaque, Pustule Bulla, Patch Nodule, Ulcer Crust, Erosion Atrophy, Exudate Telangiectasia, Scale Scar, Friable Dome-shaped, Brown(Hyperpigmentation) White(Hypopigmentation), Purple Yellow, Black
CORDA	Enlarged Cardiomediastinum Cardiomegaly Lung Opacity Lung Lesion Edema Consolidation Pneumonia Atelectasis Pneumothorax Pleural Effusion Pleural Other
DDR	Hard exudates Haemorrhages Microaneurysms Soft exudates

Instruction
The blue-whitish veil appears as an opaque, bluish-white area on the surface of the lesion, giving it a hazy or clouded appearance.

Demonstrations
Consider the following examples:
In the <image>, the blue-whitish veil is:
A) Present
B) Absent
Choose one option. Do not provide additional information.
Answer: B) Absent.
(...)

Query
In the <image>, the blue-whitish veil is:
A) Present
B) Absent
Choose one option. Do not provide additional information.
Answer:

Figure A.6: Example concept detection prompt for the Derm7pt dataset with a binary concept ("blue-whitish veil"). The same template applies to the "regression structures" concept. The instructions for all Derm7pt concepts can be found in Figure A.8. Only 1 demonstration is shown.

]

Instruction
{A brief description of the {concept}}.

Demonstrations
Consider the following examples:
In the <image>, the {concept} is:
A) Present
B) Absent
Choose one option. Do not provide additional information.
Answer: A) Present.
(...)

Query
In the <image>, the {concept} is:
A) Present
B) Absent
Choose one option. Do not provide additional information.
Answer:

Figure A.7: Example concept detection prompt for the SkinCon, CORDA, and DDR datasets. This template can be applied to each of the 22 SkinCon concepts, 11 CORDA concepts, and 4 DDR concepts (see Table A.2). The concept descriptions can be found in Figures A.9, A.10, and A.11, respectively. Only 1 demonstration is shown.

Instruction
Consider the following useful concepts to diagnose melanoma.
The pigment network consists of intersecting brown lines forming a grid-like reticular pattern. It can be absent, typical, or atypical. A typical pigment network appears as a regular grid-like pattern on dermoscopy, consisting of thin lines that form an even mesh. The spaces (holes) between these lines are relatively uniform in size and shape. In an atypical pigment network the lines forming the network are uneven in thickness, and the holes or spaces between the lines vary in size and shape.
Streaks are lineal pigmented projections at the periphery of a melanocytic lesion and include radial streaming (lineal streaks) and pseudopods (bulbous projections). They can be absent, regular, or irregular. Regular streaks are symmetrically arranged around the periphery of the lesion, appearing consistently in both length and spacing. Irregular streaks appear as projections at the periphery of a lesion and are irregular in length, thickness, and distribution.
Dots and globules can be absent, regular, or irregular. Regular dots and globules are consistent in size and shape throughout the lesion, appearing either as small, round dots or larger, round globules. Irregular dots and globules vary widely in size and shape, with some appearing small and round, while others are larger and more irregular.
The blue-whitish veil appears as an opaque, bluish-white area on the surface of the lesion, giving it a hazy or clouded appearance.
Regression structures appear as whitish, scar-like depigmented areas within the lesion, indicating areas where the pigment cells have been destroyed or the lesion has undergone partial regression.

Demonstrations
Consider the following examples:
What is the type of skin lesion that is associated with the following dermoscopic concepts: {concepts}
Options:
A) Melanoma
B) Nevus
Choose one option. Do not provide additional information.
Answer: A) Melanoma.
(...)

Query
What is the type of skin lesion that is associated with the following dermoscopic concepts: {concepts}
Options:
A) Melanoma
B) Nevus
Choose one option. Do not provide additional information.
Answer:

Figure A.8: Example disease diagnosis prompt for the Derm7pt dataset. Only 1 demonstration is shown.

Instruction
Consider the following useful concepts to diagnose skin lesions.
A papule is a small, solid, raised lesion with a diameter less than 1 cm, typically palpable and may have a distinct border.
Plaque is a broad, elevated lesion with a diameter greater than 1 cm, often with a rough or scaly surface.
A pustule is a small, elevated lesion containing pus, appearing white or yellowish, often surrounded by erythema.
A bulla is a large, fluid-filled blister with a diameter greater than 1 cm, which can be tense and easily ruptured.
A patch is a flat, discolored area of skin with a diameter greater than 1 cm, often lighter or darker than surrounding skin.
A nodule is a solid, raised lesion deeper than a papule, typically with a diameter greater than 1 cm, and may be palpable.
An ulcer is a break in the skin or mucous membrane with a loss of epidermis and dermis, often presenting as a depressed lesion with irregular borders.
A crust is a dry, rough surface resulting from the drying of exudate or serum on the skin, often forming over an ulcer or wound.
An erosion is a superficial loss of skin that does not extend into the dermis, usually appearing as a moist, depressed area.
Atrophy is the thinning or loss of skin tissue, leading to a depressed appearance, often resulting in a fragile and wrinkled surface.
An exudate is a fluid that oozes out of a lesion or wound, which can be serous, purulent, or hemorrhagic, depending on its composition.
Telangiectasia appears as dilated, small blood vessels near the surface of the skin, often appearing as fine, red or purple lines.
Scale appears as a flake or layer of dead skin cells that may shed from the surface, often seen in conditions like psoriasis or eczema.
A scar is a mark left on the skin after the healing of a wound or injury, which may be flat, raised, or depressed compared to surrounding skin.
Friable skin corresponds to skin that easily breaks or bleeds with minimal trauma, often seen in conditions like malignancies or chronic irritation.
A dome-shaped lesion is a lesion with a rounded, elevated appearance, resembling the shape of a dome, which can be smooth or irregular.
Hyperpigmentation, or an area of the skin that appears brown, corresponds to darkened skin area due to excess melanin production, often seen in conditions like age spots or melasma.
Hypopigmentation, or an area of the skin that appears white, corresponds to lightened skin area due to reduced melanin production, which may result from conditions like vitiligo or post-inflammatory hypopigmentation.
A purple area on the skin might indicate bruising or bleeding beneath the skin, often seen in conditions like purpura or ecchymosis.
A yellow area on the skin corresponds to a colour change indicating the presence of bilirubin or lipids, often seen in conditions like jaundice or xanthomas.

Dark pigmentation often indicates the presence of melanin or necrosis, frequently observed in melanoma or necrotic tissue.
Erythema corresponds to redness of the skin caused by increased blood flow to the capillaries, commonly seen in inflammation or irritation.

Demonstrations
Consider the following examples:
What is the diagnosis that is associated with the following concepts: {concepts}
Options:
A) Benign
B) Malignant
Choose one option. Do not provide additional information.
Answer: A) Benign

Query
What is the diagnosis that is associated with the following concepts: {concepts}
Options:
A) Benign
B) Malignant
Choose one option. Do not provide additional information.
Answer:

Figure A.9: Example disease diagnosis prompt for the SkinCon dataset. Only 1 demonstration is shown.

Instruction
Consider the following useful concepts to diagnose covid-19.
An enlarged cardiomediastinum is an increase in the width of the mediastinum, which may suggest conditions like aortic aneurysm, lymphadenopathy, or heart failure.
Cardiomegaly is an abnormal enlargement of the heart, often indicating underlying conditions such as heart failure, hypertension, or cardiomyopathy.
Lung opacity corresponds to an area on the radiograph where the lung appears more opaque than normal, potentially due to fluid, consolidation, or masses within the lung tissue.
Lung lesion refers to an abnormal area within the lung that may represent a nodule, mass, or area of infiltrate. It can be benign or malignant and may appear as a localised opacity or irregularity on imaging.
Edema is the accumulation of fluid in the lung interstitium or alveoli, causing increased opacity on imaging, often indicative of conditions such as heart failure or acute respiratory distress syndrome (ARDS).
Consolidation corresponds to an area of the lung where normal air-filled spaces have been replaced with fluid, cells, or other material, often seen in pneumonia or other infections.
Pneumonia is an infection of the lung parenchyma causing inflammation and consolidation, which appears as areas of opacity on radiographs.
Atelectasis is the partial or complete collapse of a part of the lung, leading to volume loss and increased opacity on radiographs, which can result from airway obstruction, compression, or surfactant deficiency.
A pneumothorax corresponds to the presence of air in the pleural space, leading to lung collapse, often visible as a dark area on the radiograph where the lung is not present.
A pleural effusion corresponds to the accumulation of fluid in the pleural space between the lung and chest wall, appearing as a blunting of the costophrenic angles or a meniscus sign on radiographs.
Pleural other refers to pleural abnormalities not classified as effusion or pneumothorax, including conditions such as pleural thickening, plaques, or calcifications.

Demonstrations
Consider the following examples:
What is the diagnosis that is associated with the following concepts: {concepts}
Options:
A) COVID-19
B) No COVID-19
Choose one option. Do not provide additional information.
Answer: A) COVID-19
(...)

Query
What is the diagnosis that is associated with the following concepts: {concepts}
Options:
A) COVID-19
B) No COVID-19
Choose one option. Do not provide additional information.
Answer:

Figure A.10: Example disease diagnosis prompt for the CORDA dataset. Only 1 demonstration is shown.

Instruction
Consider the following useful concepts to diagnose diabetic retinopathy.
Hard exudates typically appear as yellowish or whitish, well-defined spots or patches on the retina. They can vary in size and shape and often form clusters or rings around areas of retinal edema (swelling).
Soft exudates, also known as cotton-wool spots appear as pale, fluffy, or cloud-like spots on the retina. Unlike hard exudates, they are not well-defined and have a more feathery edge.
Haemorrhages can appear as red or dark spots, streaks, or blotches on the retina.
Microaneurysms appear as tiny, red dots on the retina. They are typically round and uniform in shape. They can be distinguished from other retinal features by their small size and bright red colour.

Demonstrations
Consider the following examples:
What is the type of diabetic retinopathy that is associated with the following concepts: {concepts}
Options:
A) No diabetic retinopathy
B) Mild diabetic retinopathy
C) Moderate diabetic retinopathy
D) Severe diabetic retinopathy
E) Proliferative diabetic retinopathy
Choose one option. Do not provide additional information.
Answer: B) Mild diabetic retinopathy
(...)

Query
What is the type of diabetic retinopathy that is associated with the following concepts: {concepts}
Options:
A) No diabetic retinopathy
B) Mild diabetic retinopathy
C) Moderate diabetic retinopathy
D) Severe diabetic retinopathy
E) Proliferative diabetic retinopathy
Choose one option. Do not provide additional information.
Answer:

Figure A.11: Example disease diagnosis prompt for the DDR dataset. Only 1 demonstration is shown.

Appendix B

Vision-Language Explanations

B.1 Implementation Details

The translation network f uses the standard Transformer architecture [46] with 12 layers. Across all experiments, we use 10 learnable tokens for the prefix p, as per the ablation study done by Mokady*et al.* [250], since we also did not observe an improvement beyond this number. For models trained with token and feature dropout, the dropout probability is 0.5. Since the input to f differs significantly between inference and training, a high dropout rate is beneficial. We further ensure that during training we never dropout all spatial locations or all layers for a given input, and re-sample the dropout mask in such cases.

For all experiments we use the AdamW [272] optimizer, with learning rate $= 1e-4$, weight decay $= 1e-6$, and a linearly decaying scheduler with a minimum learning rate of $1e-6$ and $5k$ warmup steps. The batch size is 64 and we limit generation to 50 tokens (20 tokens for MILANNOTATIONS experiments). Since the OPT-125M [259] language model works best on the captioning task, we use it for all other experiments and qualitative results. In our ablation with GPT-2 [184], where we strive to directly compare against ClipCap [250], we make the following changes to our translation network: lower the number of layers and attention heads from 12 to 8, and use an intermediate size of 1532 instead of the original 3072. This way we match the network hyperparameters of ClipCap. Our network ends up having fewer parameters (40M vs. 43M), since ClipCap uses a large linear layer to expand the CLIP embedding to 10 tokens while we only use 3 tokens, one per explained layer. Thus, we also achieve 35%-40% lower run time than ClipCap.

For the ResNet50 vision backbones, we explain the activations of the last layer of each residual block. The reported explanations, e.g., for layer L49, L39, L21, indicate the layer number counting all convolutional and linear layers, but correspond to a residual block each. For the ViT model, we explain each transformer layer after its feedforward projection, i.e., CLIP-ViT consists of a total of 12 transformer layers. For the MILANNOTATIONS experiments, we use CLIP-ResNet50 as well as the same ResNet101 vision backbone and explanation layers as in [33], which also use the activations after each residual block.

B.2 Limitations and Failure Cases

We use the CC3M [28] dataset for training our *DeViL* model. Even though the data covers a large variety of concepts, some words or phrases can be missing from the text data. Our model struggles to recognize more specific class descriptions, such as dog breeds like "American Staffordshire Terrier", as they appear less frequently in the data. However, it can easily recognize more general classes, like "dog", as shown in Figure B.1. Since CLIP-RN50 has a 44.45% confidence of the image showing an "American Staffordshire Terrier" when evaluating CLIP's zero-shot classification over the imagenet classes, we can conclude that this is indeed a failure case of our explanation model and not of the vision backbone. We also observed that the saliency of the model's predictions can differ for synonyms. For instance, when we compared "aircraft" and "airplane" in the second image, the saliency for "aircraft" was highly localised on the object, while it was not the case for "airplane". However, our model correctly identifies "airplane" in other images, e.g., in Figure B.5. Furthermore, when the vision backbone confuses similar features the same can be observed in our model's explanations. In the last row of Figure B.1, the region of the animal activates highly for both "horse" and "cow". When we directly contrast "cow" and "horse" in zero-shot classification, CLIP-RN50 recognizes "horse" with a confidence of 66.65% and cow with 33.35% for the given input image. Exploring these limitations, we think that training our model on more diverse and larger data can be a potential solution to address this issue, thereby improving the reliability of the model.

B.3 Feature Descriptions at Different Locations

We present additional qualitative results to showcase the efficacy of our proposed model *DeViL*. It is not only capable of explaining the different layers of the vision model being inspected but maintains the spatial arrangement of features providing meaningful descriptions at different locations of the feature map. We test our pipeline on different vision backbones and datasets. In this section, we present an analysis for CLIP-ResNet50 and CLIP-ViT-B-32. We show different layer outputs at two different locations using green and pink visual markers on the image. In Figure B.2, our model is able to detect a "windmill" (marked in green) and an "aircraft" (marked in pink). It is also able to generate meaningful captions for the lower layers of the vision backbone. For example, for the third image, the presence of flowers is noted on the final layer (L49), and in the previous layer (L39) the pink and white colours of the flowers are recognised. Similarly, we also present the output for a variety of objects present in the COCO [30] using CLIP-ViT and ImageNet [31] using CLIP-RN50 in Figures B.3 and B.4 respectively.

B.4 Saliency Maps Across Vision Backbones

In this work, we not only work on natural language descriptions but also produce open-vocabulary saliency maps corresponding to overall features from a specific layer of interest. In Figure B.5,

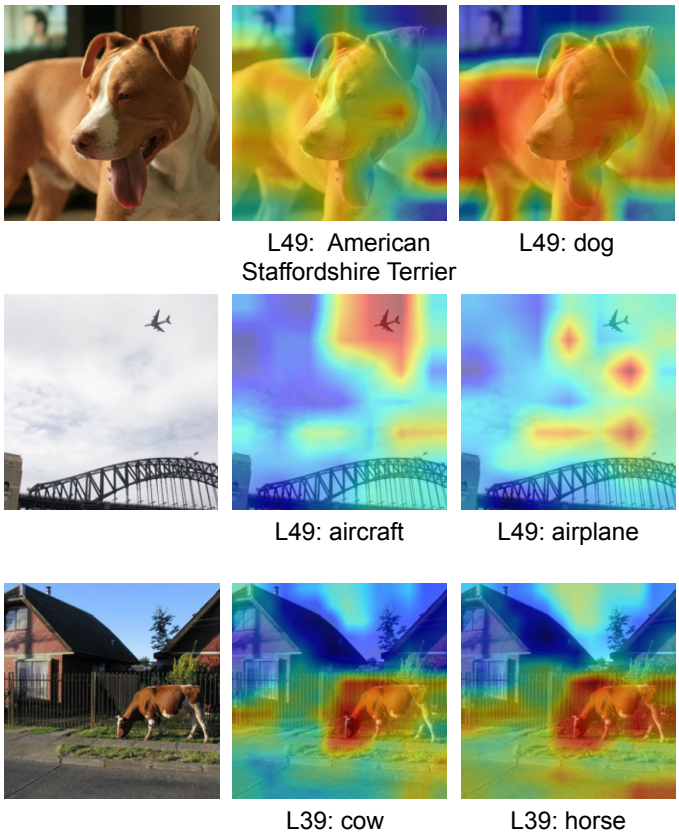


Figure B.1: Failure cases of *DeViL* for CLIP-RN50. Examples include the inability to identify specific dog breeds and specific words such as "airplane". We hypothesize that these limitations might be inherited from the limitations of the training data vocabulary of CC3M.

we present a variety of words being recognised for different scenarios. We compare different vision backbones including ResNet50 pretrained on ImageNet (IN-RN50), CLIP-ResNet50 (CLIP-RN50), and CLIP-ViT-B-32 (CLIP-ViT). We can observe that the CLIP-ViT model is able to detect objects such as "pineapple" and "boat" in the penultimate layer. While CLIP-RN50 and RN50 recognize most of the fine-grained objects in the last layer including "tent" and "elephant".

B.5 Detailed Comparison with MILAN

Figure B.6 presents additional qualitative results on the MILANNOTATIONS dataset and a comparison with the MILAN [33] model. As seen by the different generalization experiments, *DeViL* is able to generalize well and perform better than MILAN. Our model is able to correctly describe neurons using diverse language for models trained across new architectures, new datasets and training objectives. This is evident in the "Generalization across architecture" block in Figure B.6. It shows a peculiar case where *DeViL* is able to identify "people and stadiums", which is a failure case reported by Hernandez et al [33]. Our method also produces competitive results for the "Other Examples" block, where the caption produced for the leftmost case is "hands and

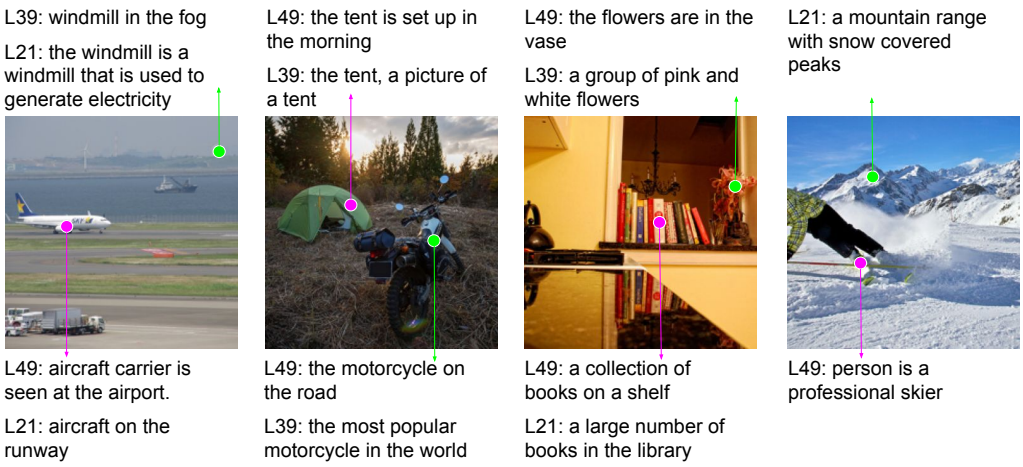


Figure B.2: Additional examples of generated localised descriptions for the CLIP-ResNet50 vision backbone on images from the COCO dataset. For each image we show two locations being explained (indicated by the pink and green markers). For each location, we present explanations from different layers of model being explained (L49, L39, etc), where a higher number corresponds to a layer closer to the output of the network.

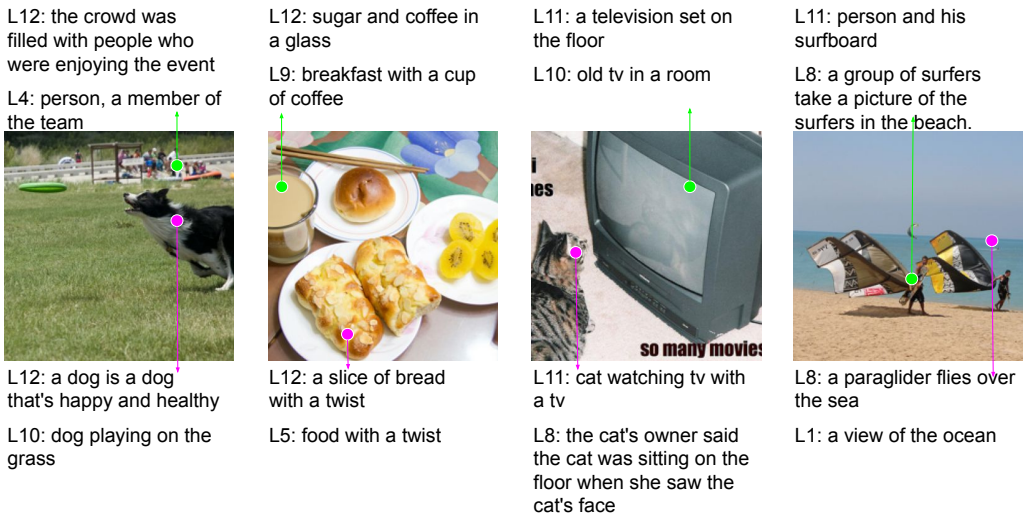


Figure B.3: Additional examples of generated localised descriptions for the CLIP-ViT vision backbone on images from the COCO dataset. For each image we show two locations being explained (indicated by the pink and green markers). For each location, we present explanations from different layers of model being explained (L12, L5, etc), where a higher number corresponds to a layer closer to the output of the network.

arms". Unlike human annotations and MILAN, *DeViL* not only mentions hands, but also arms, which are also visible in the masked images.

We also compare quantitatively with the state-of-the-art network dissection MILAN model [33] and with ClipCap [250]. Although our model was designed to work with layer-wise feature maps for a single image and not at the neuron level, we adapt *DeViL* by pooling over the 15 masked

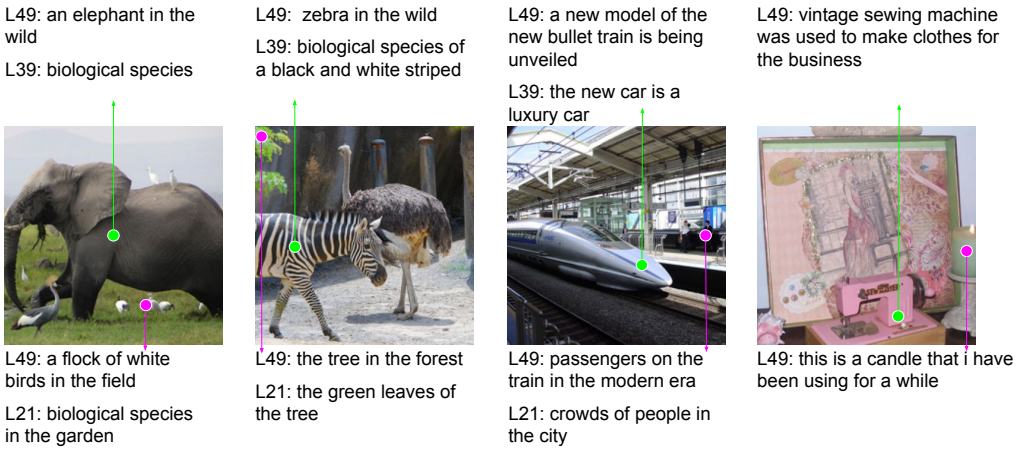


Figure B.4: Additional examples of generated localised descriptions for the CLIP-ResNet50 vision backbone on images from the ImageNet dataset. For each image we show two locations being explained (indicated by the pink and green markers). For each location, we present explanations from different layers of model being explained (L49, L21, etc), where a higher number corresponds to a layer closer to the output of the network.

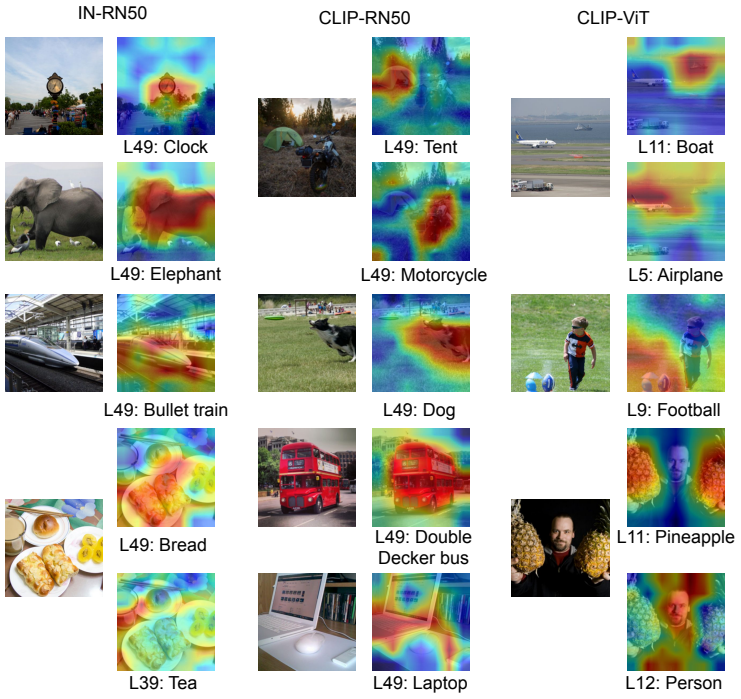


Figure B.5: Additional open-vocabulary saliency examples for different vision backbones (IN-RN50: ResNet50 pretrained on ImageNet, CLIP-RN50, and CLIP-ViT). For each example, we present the original image alongside saliency maps generated for different words and different layers of the models being explained (L49, L11, etc). A higher layer number corresponds to a layer closer to the output of the network.

images given for each neuron. We report the BERTScore [211] results in Table B.1 (higher is better), on several of the generalization experiments proposed by the MILAN [33] authors. For fair comparison, both ClipCap and *DeViL* use CLIP-RN50 as the vision backbone in Table 5.3. In Table B.1, we additionally report the results using ResNet101 (RN101) as vision backbone for *DeViL* as it was used by MILAN [33]. We are able to outperform both MILAN [33] and ClipCap [250] on mean BERTScore with either backbone.

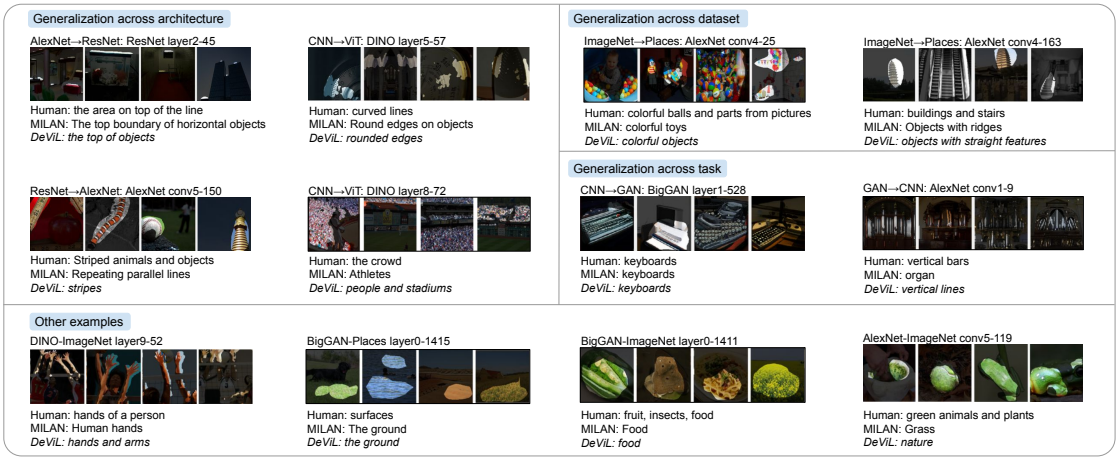


Figure B.6: Additional qualitative examples on the MILANNOTATIONS dataset and comparison with the MILAN [33] model. For each image in the generalization blocks, we specify the generalization experiment (e.g., AlexNet→ResNet) before the model, layer and neuron being explained.

Table B.1: Results on generalization experiments with the MILANNOTATIONS [33] dataset, in terms of BERTScore (higher is better).

Generalization	Train	Test	MILAN [33]	ClipCap [250]	*DeViL* (RN101)	*DeViL* (CLIP-RN50)
Within network	AlexNet-ImageNet		0.39	**0.43**	**0.43**	**0.43**
	AlexNet-Places		**0.47**	**0.47**	0.46	0.46
	ResNet152-ImageNet		0.35	**0.38**	**0.38**	**0.38**
	ResNet152-Places		0.28	0.33	**0.37**	0.33
	BigGAN-ImageNet		0.49	**0.51**	0.49	**0.51**
	BigGAN-Places		0.52	0.50	0.53	**0.54**
Across arch.	AlexNet	ResNet152	0.28	**0.33**	0.30	0.30
	Resnet152	AlexNet	0.35	0.35	**0.36**	**0.36**
	CNNs	ViT	0.34	**0.38**	0.37	0.37
Across datasets	ImageNet	Places	0.30	0.32	**0.33**	**0.33**
	Places	ImageNet	**0.33**	0.31	**0.33**	0.32
Across tasks	Classifiers	BigGAN	0.34	**0.37**	0.36	0.35
	BigGAN	Classifiers	0.27	0.27	0.28	**0.29**
Mean			0.362	0.381	**0.384**	0.382