

---

SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science  
Department of Computer Science  
MASTER THESIS

---



# Interpretability in Vision-Language Models via Cross-Modal Alignment under the Concept Bottleneck Model Framework

submitted by  
Zurana Mehrin Ruhi  
Saarbrücken  
May 2025

---

**Advisor:**

Md Abdul Kadir  
German Research Center for Artificial Intelligence  
Saarland Informatics Campus  
Saarbrücken, Germany

**Reviewers:**

Prof. Dr. Daniel Sonntag  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

Prof. Dr. Antonio Krüger  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

**Submitted**

26th May, 2025

Saarland University  
Faculty MI – Mathematics and Computer Science  
Department of Computer Science  
Campus - Building E1.1  
66123 Saarbrücken  
Germany

## Acknowledgements

I would like to express my sincere gratitude to my advisor, **Md Abdul Kadir**, for his guidance, patience during our long discussions, and iterative feedback on implementing various approaches. His assistance guided me in shaping this research and bringing this thesis to completion.

I also extend my gratitude and utmost appreciation to **Prof. Daniel Sonntag** for his valuable feedback and work at the forefront of this domain, which inspired me to continue this research.

I would also like to thank my friend, Sigma Jahan, for her motivational words and for sharing her research experiences.

I am highly grateful to my sister, Dr. Jannatul Mehjabin Juhy, and my brother, Dr. Mahy Md Murtayes Jubayer, for supporting me throughout this research with their clinical expertise.

Lastly, I thank my parents Parvin Begum and Dr. Md Abu Zaher for their continuous support.

## Abstract

Vision-language models are powerful tools for versatile and interactive applications given their capability to leverage multimodal data. However, integrating these models in real-world, safety-critical domains remain challenging due to their limited interpretability, and difficulties in ensuring robust performance for unseen scenarios. This thesis addresses the challenge of improving the interpretability of vision-language foundation models, particularly in understanding the correlation between image regions and their associated textual descriptions. The primary aim is to develop a practical methodology for mapping which specific visual elements directly influence particular text outputs. This is particularly important for applications such as automated radiology report generation, where accurately correlating image findings to text can significantly enhance the clarity and reliability of reports and allow for safety checks in novel cases. Utilizing techniques from explainable artificial intelligence (XAI), this research aims to (a) identify visual concepts that significantly influence text generation, (b) systematically extract human-understandable visual cues and (c) establish a data-driven link between textual and visual concepts contributing to the final output. Such cross-modal concept-based explanations can help the user comprehend how vision-language models process and utilize information across different modalities. To achieve this, we propose a concept-bottleneck-based approach; which shows promising results on selected datasets despite not generalizing reliably across all scenarios. In this work, we discuss both the potential and limitations of concept-based interpretability by providing deeper insights into the model's decision-making process and the practical challenges of cross-modal alignment. This work seeks to contribute and motivate future work towards more transparent and interpretable architectures, thereby increasing user trust in using AI-assisted applications.



---

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Motivation . . . . .	3
1.2	Problem Definition . . . . .	3
1.3	Research Objectives . . . . .	4
1.4	Scope and Contribution . . . . .	4
<b>2</b>	<b>Related Work</b>	<b>6</b>
2.1	Vision-Language Models . . . . .	6
2.2	Explainable AI techniques . . . . .	8
2.2.1	Explanation by Feature . . . . .	9
2.2.2	Interpretation by Concept . . . . .	9
2.3	Inherently Interpretable AI . . . . .	10
2.3.1	Concept Bottleneck Models . . . . .	11
2.3.2	CBMs for Vision-Language Models . . . . .	12
2.4	Domain-specific Approaches: Medical AI . . . . .	13
2.5	Summary . . . . .	15
<b>3</b>	<b>Technical Background</b>	<b>17</b>
3.1	Transformers . . . . .	17
3.2	Vision Transformers (ViTs) . . . . .	18
3.3	Vision-Language Models (VLMs) . . . . .	18
3.4	XAI for Vision-Language Models . . . . .	20
3.4.1	Concept Bottleneck Models for VLMs . . . . .	20
3.5	Evaluation Metrics . . . . .	21
<b>4</b>	<b>Methodology</b>	<b>23</b>
4.1	Visual Concept Generation . . . . .	24
4.1.1	Concept Discovery via Segmentation . . . . .	25
4.1.2	Per-Input Equivalence (PIE) . . . . .	25
4.1.3	Shapley-Based Attribution . . . . .	26
4.2	Textual Concept Extraction via Prompting . . . . .	27
4.3	Concept Bottleneck Modeling (CBM) . . . . .	28

4.3.1	Architecture Overview . . . . .	28
4.3.2	Visual-Textual Concept Supervision . . . . .	28
4.4	Design Considerations . . . . .	31
<b>5</b>	<b>Experiments and Evaluation</b>	<b>33</b>
5.1	Dataset Description . . . . .	33
5.2	Experimental Setup . . . . .	34
5.3	Evaluation of Visual Concept Generation . . . . .	35
5.3.1	Impact of Domain-Specific Models . . . . .	36
5.3.2	Performance Across Medical Datasets . . . . .	36
5.3.3	Qualitative Interpretability . . . . .	37
5.4	Evaluation of Concept Bottleneck Model (CBM) . . . . .	41
5.4.1	Evaluation on MIMIC-CXR . . . . .	41
5.4.2	Evaluation on CheXpert . . . . .	42
5.4.3	Qualitative Interpretability . . . . .	42
5.4.4	Global Concept Attribution . . . . .	45
5.4.5	Visual-Textual Concept Alignment . . . . .	48
5.5	Interpretability and Utility of Generated Explanations . . . . .	49
5.6	Case Studies and Limitations . . . . .	51
5.6.1	Case 1: Lung Opacity . . . . .	52
5.6.2	Case 2: Cardiomegaly . . . . .	53
5.6.3	Limitations . . . . .	53
<b>6</b>	<b>Conclusion</b>	<b>56</b>
6.1	Future Work . . . . .	57
6.1.1	Integration with Large Language Models (LLMs) . . . . .	57
6.1.2	Spatially-Aware and Adaptive CBMs . . . . .	57
6.1.3	Causal and Hierarchical Concept Modeling . . . . .	58
6.1.4	Human-in-the-Loop and Editable CBMs . . . . .	58
	<b>Bibliography</b>	<b>59</b>

---

# List of Figures

2.1	Primary components in Vision-language models: Decoder-only Transformer (a) architecture is used for language understanding and Vision Transformers (b) for image understanding. . . . .	7
2.2	Contrastive Language Image Pre-training Architecture [98] that uses a contrastive learning objective to align image and text embeddings. . . . .	7
2.3	General high-level ontology of Explainable AI approaches. While transparent models exist which are easily interpretable, most models are black-box and categorized as opaque. To make these methods explainable, model-specific or various model-agnostic approaches are applied [9]. . . . .	8
2.4	Concept Relevance Propagation proposed the idea of conditional back-propagation linked to a specific concept or latent representations of multiple concepts within or across layers [2]. . . . .	10
2.5	Recently developed methods that use segmentation techniques to support and improve performance of existing methods [68, 118]. . . . .	11
2.6	Visual Concept Filtering using activation scores in CBMs [66]. Instead of manually creating a concept set, an LLM is used to generate concepts per disease. Then concepts are filtered via submodular optimization (which is a mathematical approach applied to a set of concepts and acts similar to feature selection techniques) and concepts with the highest visual activation scores are selected to train the final bottleneck layer. . . . .	12
2.7	Label-free Concept Bottleneck Models [91] automate the process of text concept creation using GPT models. A CBL layer is then learned with the objective of maximizing similarity between the concept activation vector and the concept matrix, which is the joint representation of projection of image and concept embeddings. This helps the model focus on the most meaningful concepts by maximizing the alignment between the activations and the image-derived concepts. The final layer then minimizes the classification error with the objective to improve accuracy while maintaining interpretability through a sparse representation. . . . .	13
2.8	Concept Hierarchy [16] defined in two levels of granularity and compared to image patches using CLIP's contrastive learning approach. . . . .	14
2.9	Grad-CAM based explanations for Medical Chest X-rays [87] . . . . .	14
2.10	Concept based explanations for Medical Chest X-rays [100] . . . . .	15
4.1	Overview of the proposed visual-language concept bottleneck framework. Visual concepts are extracted via a trained and interpretable concept generator module. A concept bottleneck layer, trained with a sparse classifier, maps these concepts to the final task output. The framework also supports influence analysis to quantify visual and textual contributions. . . . .	23
4.2	Overview of the proposed Visual concept generation framework. . . . .	25

5.1	Extracted visual concept attributions on frontal chest X-rays. Their SHAP values overlaid on segmented regions reflect their importance in predicting respective findings. . . . .	38
5.2	Visual Concept based interpretability across different viewpoints and findings. . . . .	39
5.3	Top 5 concepts for each class across different Chest X-ray Samples from COVID-QU [29], MIMIC-CXR [56] and CheXpert [21]. In this approach, concept discovery is highly dependant on the segmentation model used, thus concepts may lack completeness in clinical semantics. The average accuracy across all datasets is 83.42%; indicating the potential of visually interpretable classification using this approach and therefore concept discovery can be highly improved by using a more accurate medical segmentation model. . . . .	40
5.4	Visual and Textual concept attributions on Sample chest X-rays. The SHAP values overlaid on segmented regions reflect their importance in predicting respective findings. The Concept scores indicate the weight of the learned sparse layer in CBM architecture. . . . .	44
5.5	Learned concept contributions for <b>Cardiomegaly</b> . Positive contributions are shown in blue, negative in red. Weights are taken from the sparse linear classifier in the CBM. . . . .	45
5.6	Concept contributions to <b>Lung Opacity</b> . CBM captures fine-grained inter-concept effects including suppression by clear findings. . . . .	46
5.7	Learned concept contributions to <b>COVID-19</b> . Findings like “Left lung cavitation” and “Normal mediastinal contour” contribute positively. . . .	46
5.8	Concept-level decision patterns for <b>Non-COVID</b> and <b>Normal</b> classes. The model learns to suppress abnormal features in the Normal class. . . . .	47
5.9	Top-5 textual concepts semantically aligned with visual segments using cosine similarity. These associations confirm the medical interpretability of extracted visual concepts. . . . .	48
5.10	Visually segmented concepts contribute towards higher interpretability and quick understanding of model decisions based on important regions. In lateral views, when model is confused or misidentifies regions, visual importances make it easily detectable. . . . .	50
5.11	Concept contributions for COVID and Non-COVID disease diagnosis is distinct, aligning with clinically important factors such as pleural effusion being an indicator of the presence of COVID infection. . . . .	51
5.12	Explanation of the model’s prediction for the <b>Lung Opacity</b> class (Confidence: 0.593). . . . .	52
5.13	Explanation of the model’s prediction for the <b>Cardiomegaly</b> class (Confidence: 0.598). . . . .	53
5.14	Limitations of Visual Segments and failure cases. . . . .	54

## List of Tables

5.1	Total sample counts for each finding across MIMIC-CXR, CheXpert, and ChestX-ray8 datasets . . . . .	34
5.2	Mean AUC (%) and standard deviation (%) for visual concept attribution across 5 runs. . . . .	36
5.3	Performance impact of MedSAM on CheXpert (5 runs). . . . .	36
5.4	Faithfulness - AUC scores with standard deviation across medical datasets (5 runs). . . . .	36
5.5	Comparison of concept bottleneck models in medical imaging. Our method uses no manual labels and achieves competitive accuracy. . . . .	41
5.6	Ablation on modality combinations, similarity cutoff, and sparsity vs. performance tradeoffs for MIMIC. All models use CLIP ViT-B/32 as backbone. . . . .	41
5.7	CBM results on CheXpert . All models use max-pooling for visual concept aggregation. . . . .	42

---

# Chapter 1

## Introduction

Artificial intelligence (AI) has been continuously proving its potential to transform healthcare, particularly through the use of recent state-of-the-art models for tasks such as diagnosis support [73], medical image analysis [58], and clinical report generation. Among these, use of vision-language models (VLMs) are a key advancement due to their ability to combine visual and textual modalities, allowing AI-enabled applications that can investigate medical images and textual findings in human like nature. This multimodal reasoning capability offers the possibility of more comprehensive and context-aware Medical AI systems. However, despite their progress on benchmark datasets, significant challenges remain for safely deploying such model in healthcare environments. One of the most critical issue among these challenges is the difficulty in interpreting how such models correlate visual features with corresponding textual outputs. In the context of such critical and life-dependant decision-making, such as radiological diagnosis or report generation, transparency in model reasoning is the key to user trust. Therefore, the higher complexity and lack of interpretability highly limits the adoption of AI assistance into clinical workflows[72].

Although Explainable AI domain is a well-researched domain providing valuable insights into model behaviour, the techniques applied to vision-language models in complex, real-world domains are still not established. Methods such as saliency maps [110] and attention visualizations offer limited interpretability [2], and struggle when reasoning across modalities is involved. Moreover, there remains a gap between post-hoc explanations and the need for human-understandable explanation that can support by medical professionals.

Addressing these limitations requires a deeper investigation into how visual and textual concepts interact within VLMs and how they can be systematically aligned in a clinically meaningful way. In this context, the following section articulates the motivation for this thesis, focusing on the need for structured, concept-based interpretability methods that can enhance transparency, support clinical validation, and ultimately foster trust in AI-assisted medical decision-making.

## 1.1 Motivation

Despite the success of VLMs both in visual-question answering and natural language understanding, the deployment of these models in safety-critical fields poses a very high-staked challenge due to their limited interpretability and vulnerability to unseen scenarios [53, 20]. In clinical practice, where model decisions can significantly impact diagnosis and treatment outcomes, understanding the reasoning behind AI predictions in such clinical decision support systems is critical along with safety checks when a model might make wrong predictions; for ensuring transparency, trust, and accountability[71].

In medical imaging, particularly radiology, explainable artificial intelligence (XAI) approaches at its early stages only relied on post-hoc techniques such as saliency maps [114, 110]. However, these methods frequently produce coarse, ambiguous, or even misleading explanations [2], reducing their utility in practical decision-making applications [53]. Moreover, vision-language models introduce additional complexity by operating across different modalities, making it even more difficult to explain which specific visual elements influence their textual counterparts. Without any mappings or understanding between visual regions and generated reports, medical professionals remain skeptical of AI-generated results [128], along with regulatory compliance under frameworks such as GDPR and HIPAA becoming stricter.

Recent efforts in concept-based interpretability offer a meaningful direction of understanding the decision-making of such models by projecting latent representations into semantically meaningful subspaces [137, 91]. However, most existing concept bottleneck models are designed for unimodal settings and require manual annotation to concept curation, limiting their scalability to real-world scenarios. Particularly in medical applications, where data curation and annotations are dependant on expert opinion, there is a demand for automatic extraction and validation of concepts in a human-understandable way [20].

This thesis aims to bridge this critical gap by developing a cross-modal, data-driven methodology that systematically maps image regions and textual descriptions in an interpretable manner. By leveraging principles of XAI and integrating visual concept extraction with textual concept alignment, the proposed approach seeks to offer both local (instance-specific) and global (class-level) interpretability. Such a framework is essential to allow safety checks, auditing, and build clinician’s trust in AI-assisted decision support systems. Through transparency and conceptual traceability, this research delved into explaining vision-language models from opaque black-box systems towards clinically viable, interpretable, and trustworthy solutions.

## 1.2 Problem Definition

In currently established research works, there is a lack of specificity in how image regions contribute to particular elements of the textual descriptors [132]. While attention mechanisms and saliency maps attempt to offer visual explanations, they often produce coarse and clinically ambiguous outputs that are insufficient for supporting medical decision-making [87]. Without understanding the connections between visual findings and textual descriptions, these models risk propagating errors, generating hallucinated content, or missing clinically significant features in unseen scenarios.

Therefore, there is a critical need for methodologies that can (i) extract human-interpretable

visual concepts from medical images without relying on extensive manual annotation, (ii) systematically associate these visual concepts with corresponding textual outputs, and (iii) provide transparent, faithful explanations of the model’s decision process. Addressing this problem would not only improve the interpretability of vision-language systems but also help develop a way towards more trustworthy AI applications in clinical radiology.

### 1.3 Research Objectives

The goal of this research is to contribute toward improving the interpretability of vision-language models in the context of medical imaging, with an emphasis on identifying both visual and textual descriptors. This thesis is structured around the following objectives:

- To design a systematic approach for extracting semantically meaningful visual concepts from radiographic images, minimizing reliance on manual annotation.
- To develop methods for associating extracted visual concepts with clinically relevant textual concepts generated by vision-language models, enabling traceability between input features and model outputs.
- To implement a concept-based bottleneck architecture that facilitates interpretable intermediate representations that encourages consistency between visual concepts and textual descriptions during training.
- To evaluate the proposed methodology using publicly available medical imaging datasets, assessing both predictive performance and interpretability through quantitative metrics and qualitative analysis.

These objectives are intended to address the challenges outlined in the problem definition by providing structured mechanisms for aligning visual and textual reasoning within multimodal models, with the aim of supporting safer and more transparent clinical applications.

### 1.4 Scope and Contribution

This thesis focuses on the interpretability of vision-language models applied to radiology imaging, with a specific emphasis on understanding and explaining the associations between visual regions and corresponding textual outputs. The scope is within the domain of radiology imaging, and the scope is limited to tasks involving chest X-ray datasets and related clinical concepts.

The contributions of this research are as follows:

- **Visual Concept Extraction:** A method is proposed to extract visual concepts from radiographic images, aiming to identify regions that are semantically meaningful without requiring extensive manual annotations.
- **Cross-Modal Concept Alignment:** Techniques are developed to systematically associate visual concepts with corresponding textual concepts generated by vision-language models.



- **Concept Bottleneck Layer Integration:** A concept bottleneck architecture is incorporated into the model pipeline, providing an interpretable intermediate representation that captures critical information before making predictions.
- **Custom Loss Function Formulation:** A tailored loss function is designed to encourage alignment between visual features and textual outputs during model training, supporting more coherent and interpretable reasoning.
- **Evaluation on Public Datasets:** The proposed methods are implemented and evaluated on publicly available datasets, with assessments based on both predictive performance and the interpretability of generated explanations.

The scope of the study does not extend to explaining or evaluating language generation itself but focuses on establishing connection between pre-generated concepts. Due to computational resource limitations, evaluating the work on advanced VLMs is out of scope.

The remainder of this thesis is structured as follows. Chapter 2 provides a detailed review of related work, covering vision-language models, explainable AI techniques and domain-specific approaches in medical AI. Chapter 3 presents the necessary technical background, including an overview of transformers, vision transformers, vision-language models, and explainability methods relevant to this study. Chapter 4 describes the methodology developed for visual concept extraction, textual concept generation, integration of a concept bottleneck layer, and formulation of custom training objectives. Chapter 5 discusses the experimental setup, including dataset processing, model selection, evaluation procedures, and analysis of results on classification and visual question answering tasks. Finally, Chapter 6 concludes the thesis by summarizing the findings, discussing limitations, and outlining directions for future research.

---

## Chapter 2

### Related Work

Vision-Language models are powerful architectures that can learn to perform a variety of tasks such as image captioning, visual-question-answering and even cross-modal generation. Such AI models are more contextually aware than classical models. However, the more complex a model gets, the less interpretable and explainable it becomes, making it less trustworthy for using in critical domains. Therefore, introducing interpretability can help to improve human interaction with these complex models. In this report, after a brief overview of the architecture of Vision-language models, how these models or their individual components can be made human-understandable is explored. The literature work is divided into two sub-sections:

- Explainable AI approaches: primarily post-hoc methods are explored that do not require target model training and focus on explaining how the model generated a particular decision [9].
- Inherently Interpretable AI approaches: methods that require training to make the model systematically conclude interpretable decisions are explored [9].

#### 2.1 Vision-Language Models

Vision-Language models combine both the strength of Vision Models and Large Language models to process visual and textual data within the same pipeline. The transformers [123] architecture was first introduced for Language related tasks where sequence-to-sequence modelling is essential to maintain long-term coherence [49] within a piece of text. Self-attention and multi-headed attention [31] mechanisms were introduced to capture the relationship within and between words. The same theory later was adapted for images in Vision-transformers [42], where each image is divided into patches that are treated like words in a sequence. This allows vision-transformers to capture the relationship within image patches to get a global embedding of the whole image, as opposed to the limited global understanding of Convolutional Neural Networks (CNNs) [115].

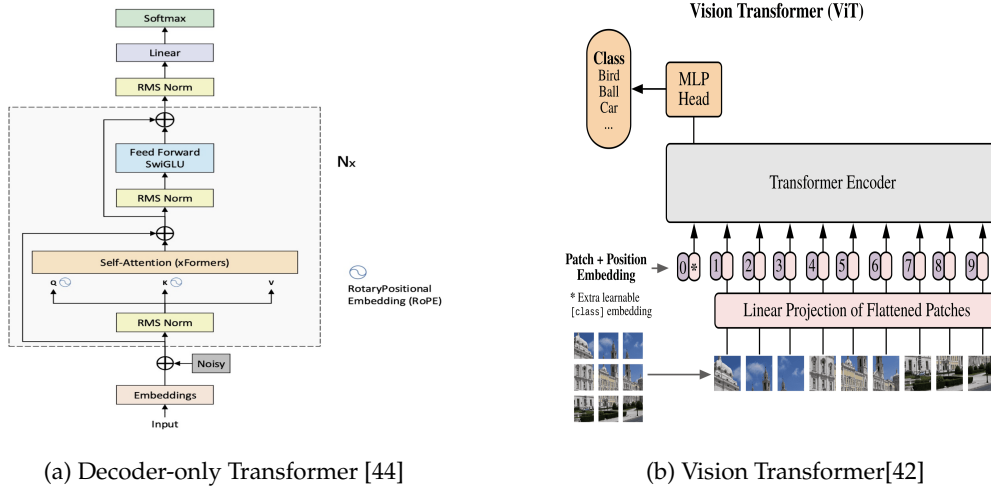


Figure 2.1: Primary components in Vision-language models: Decoder-only Transformer (a) architecture is used for language understanding and Vision Transformers (b) for image understanding.

Contrastive Language-Image Pre-training (CLIP) [98] model was created by combining these two architectures. CLIP is a multimodal neural network that can capture the relationship between its individual unimodal embedding spaces by aligning them in a shared embedding space. Many models and approaches were later introduced to take this one step further by incorporating cross-modal attention [23] mechanism. Whereas CLIP uses contrastive learning to align embeddings, cross-modal attention method is applied to learn multimodal embeddings by jointly processing both modalities within attention layers. This mechanism allows the model to capture fine-grained interaction between images and texts and has been widely researched [76, 6, 75, 93, 1, 28]. Due to Vision-Language models' broad applicability, various learning paradigms such as instruction tuning, masked language modelling, image-text matching, masked region modelling, generative training and others were introduced to make these models more efficient and adaptive [84, 25, 117, 101].

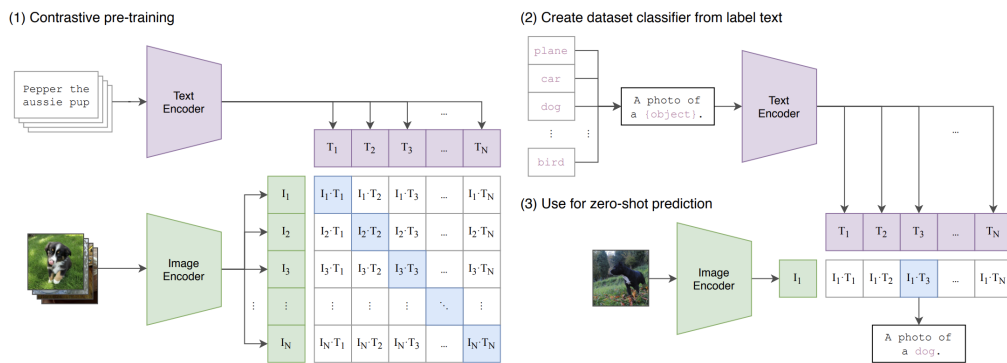


Figure 2.2: Contrastive Language Image Pre-training Architecture [98] that uses a contrastive learning objective to align image and text embeddings.

Vision-Language Foundation models, in this context are comprised of vision-language models that uses pre-training paradigm [127, 135, 79, 80, 33] and being trained on very large datasets, can perform a wide range of multimodal tasks. Domain-specific models already exist for healthcare, automotive, finance, 3D image understanding and even geospatial domains [12, 138, 15, 27, 116, 129, 54]. LVM-med [89] is a medical foundation model that is trained on various modalities of data such as MRI, X-rays, CT-scans and improves performance for medical-specific tasks. Needless to say, these models are black-box and non-interpretable, preventing their wide-spread trust-based deployment in real world.

## 2.2 Explainable AI techniques

Explainable AI is a branch of AI that seeks to explain AI models in a systematic way to establish user trust and perform sanity checks [9]. Based on training stages, we can divide the approaches in two: explain the model behaviour after training and train the model to behave in desired rules of interpretation. Before we dive deeper, it is important to understand the societal impact of such approaches in figure 2.3. Social Scientists argue that explanations should be relevant in context and should be understandable by non-technical users (that is probabilities probably do not matter) [88]. Some popular methods like Saliency maps are even found not reliable or generating random explanations [4, 121]. Therefore, while exploring such methods, it is important for a data scientist to be aware of the purpose and use of methods.

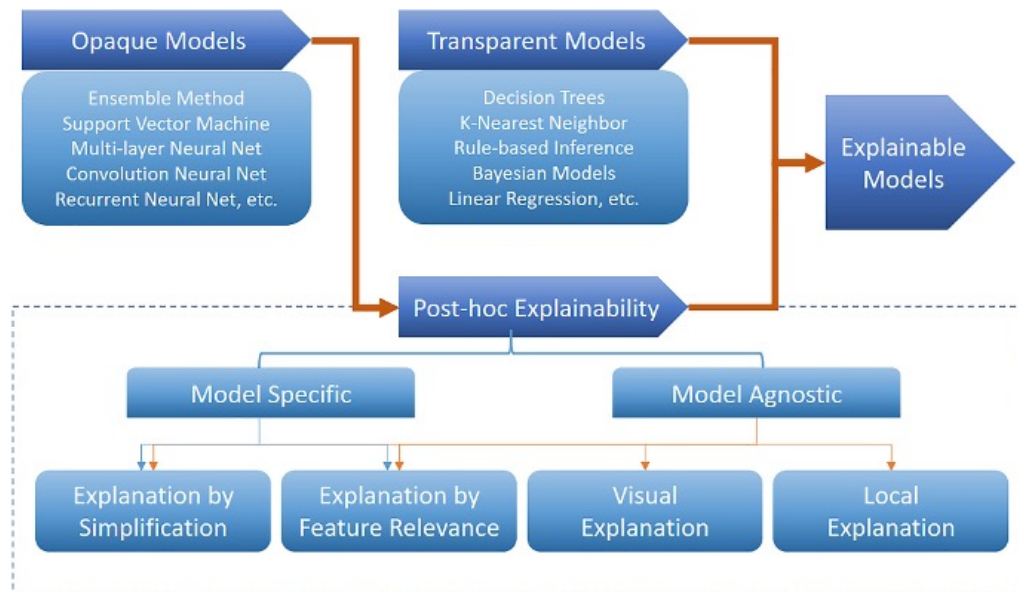


Figure 2.3: General high-level ontology of Explainable AI approaches. While transparent models exist which are easily interpretable, most models are black-box and categorized as opaque. To make these methods explainable, model-specific or various model-agnostic approaches are applied [9].

Within the context and scope of this work, we deal with following concepts of Explainable AI (XAI):

- **Explainability:** An interface between humans and AI systems that make the model accurate and comprehensible to human understanding [47].
- **Interpretability:** The model's capacity to provide interpretable decisions that are easily understandable by humans. [47].
- **Model-agnostic:** XAI approach that can be applied to any model to explain its outputs given its inputs, as opposed to model-specific approaches that leverage the knowledge of the model architecture and parameters [39, 103].
- **Explanation by Feature:** XAI approaches that rely on input's features or attributes to explain the decision-making process of the model [85, 11]
- **Interpretation by Concept:** XAI approaches that connect the decision-making process or internal workings of AI models to high-level concepts that are understandable by humans [63, 32]

### 2.2.1 Explanation by Feature

Feature Relevance techniques rank the features to explain the model's decision by conveying which feature contributed to the final output. Saliency[114] maps and gradient-based methods were formulated to visualize such features; while these methods provide somewhat explanation on what part of image the model is focusing on, reliant on back-propagation they often fail to encapsulate sufficient information on data distribution [57]. EMILE-UI [59] employs feature removal techniques while incorporating interactive explanations for improved human understanding. Shapely values [85], originated from game theory, tries to find most explanatory features by distributing importance to each feature and calculating weighted contribution of every possible combination of features on the model's output. While it is model agnostic, applying it on non-additive model may not provide useful explanations or may not even be efficient. Layer-wise Relevance Propagation (LRP) [11] on the other hand calculates a relevance score starting from the final output layer and backpropagates it through each layer of a network until it reaches the input layer. This was particularly developed for Deep Neural Networks, to explain how relevance flows through the entire network, providing local explanations. However, relevance flow provides limited interpretability and the model-specificity of LRP can make it unusable on complex architectures. Over the years, LRP has been developed for Deep Learning models [112, 13, 7, 102] showing its capacity in providing meaningful explanations. These works provide a pathway to explain specific components of vision-language models, such as generating attention-aware explanations for transformers [3] using LRP formulation or using shapely values to explain CLIP model's decision behaviour [118].

### 2.2.2 Interpretation by Concept

Concept Activation Vectors [63] were introduced to quantify the degree to which a concept influences a model's prediction, making the model interpretable. These concept-based explanations do not focus on granular data but extracting abstract and complete features across the entire dataset (i.e. higher level concepts). In this work, a concept set was pre-defined to support human understandability and flexibility. Methods exist to reduce the amount of manual intervention and risk of bias to create such concept sets [45] and even incorporate game-based approaches to rank the most meaningful concepts

[136]. Concept Relevance Propagation [2] further extended the idea by combining LRP to provide both local (concept relevance heatmaps) and global (concept relevance score) explanation for concepts, as seen in figure 2.4

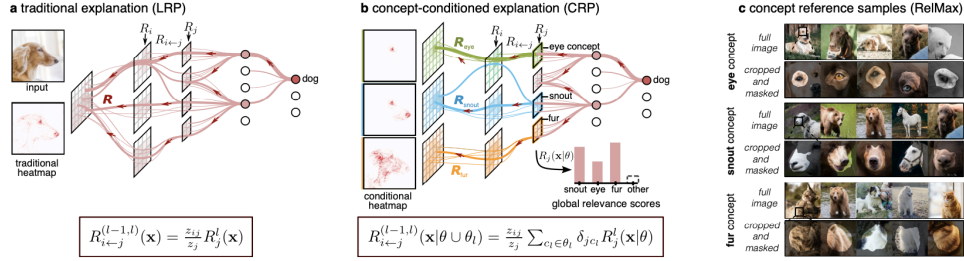


Figure 2.4: Concept Relevance Propagation proposed the idea of conditional backpropagation linked to a specific concept or latent representations of multiple concepts within or across layers [2].

Even though these can help explain and even validate model decisions using additional prototypical concepts [43], they still inherit the limitations LRP. Furthermore, how to choose the best set of concepts are still unknown to these methods and often comes pre-defined, hand-crafted or with the assumption of a mixture-of-Gaussian (MoG) distribution of concepts [43]. A concept discovery approach [124] attempts to remove the constraints of earlier method by creating multidimensional concept subspaces, however the algorithm is generic and may not be transferable to Large Language Model architectures in practice.

## 2.3 Inherently Interpretable AI

Linear regression, logistic regression and decision trees are known to be interpretable models [65] as humans can easily understand the underlying cause for their decisions. An early approach for making any given model interpretable is Local Interpretable Model-Agnostic Explanations (LIME) [104]. However, it uses perturbation methods with random sampling to generate explanations, which may provide inconsistent explanations. Furthermore, it only approximates feature contribution locally and is sensitive to artifacts. Although Anchors [105] try to mitigate some limitations of LIME by improving precision and addressing the issue of inconsistency, the limited coverage of this method and inability to generate complex rules in presence of high-dimensional features, make it unusable for complex models. However, LIME has been a cornerstone to develop improved locally interpretable methods such as ALIME [113], SLIME [130], GLIME [120] and DSEG-LIME [68] which can be applied directly to Vision-transformers to find meaningful visual concepts. Explain-Any-Concept [118], as seen in figure 2.5b, also introduces segmentation techniques [67] to create visual concepts and train a surrogate model with shared fully-connected layers that is supposed to overcome the limitations of Linear LIME models, however this may lead to unrelated concept generation irrespective of task intended.

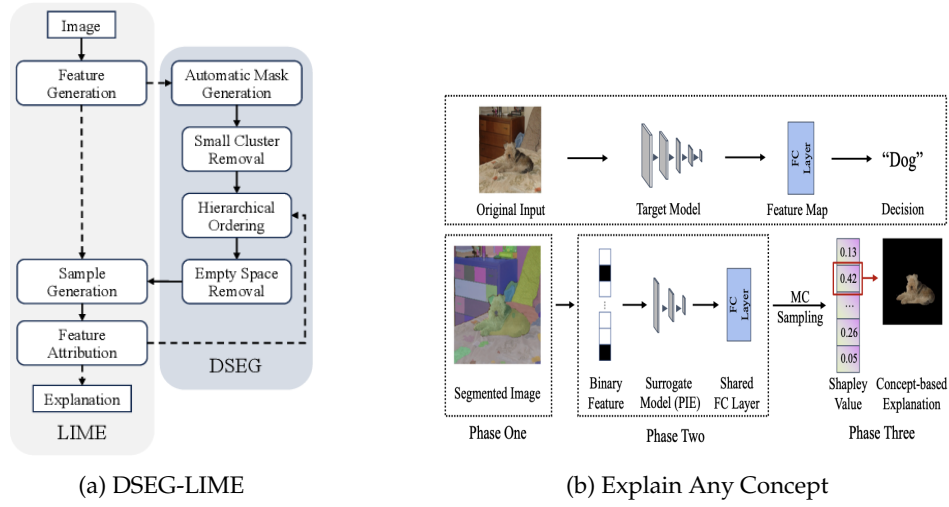


Figure 2.5: Recently developed methods that use segmentation techniques to support and improve performance of existing methods [68, 118].

### 2.3.1 Concept Bottleneck Models

Concept Bottleneck Models (CBMs) are designed to be inherently interpretable as they focus on finding a set of concepts that are human-understandable and then use these concepts to draw final decisions. It is a simple idea to introduce an intermediate layer that learns these concepts and adjusts the model’s decision boundaries based on these concepts [69]. There have been many approaches to enhance such model’s interpretability in different scenarios. Policy-based interactivity [22] even allows concept set selection for particular labels to maximize the final prediction decision. Visual concept filtering uses concept activation scores [66] to measure whether the visual cues in predicted concepts are relevant or not before learning to predict final outcomes. Probabilistic CBMs [64] additionally adds an uncertainty score using probabilistic concept embeddings to increase user trust and reduce ambiguity. Additional unsupervised concepts can also be included to improve learning of the predicted concepts [109]. Hierarchies within concepts were also introduced to capture understandable concepts on two-levels of granularity. Use of data-driven coarse-to-fine selection methods and Bayesian sparsity can make these sort of framework highly interpretable [95]. CBMs has been since applied for various models within the context of image/text classification, object detection, semi-supervised tasks and error monitoring in semi-supervised and even self-supervised concept learning settings [62, 51, 107]. However, the non-visual concept filtering via submodular optimization (as seen in step II of figure 2.6) in domain-specific image classification [132] might leave out subtle concepts that is important for original classification tasks, hence affecting model’s performance.

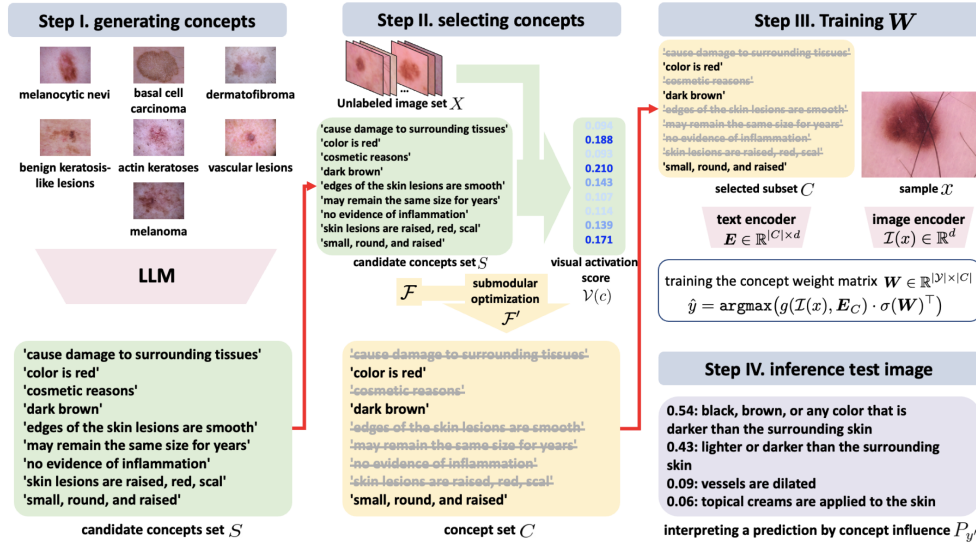


Figure 2.6: Visual Concept Filtering using activation scores in CBMs [66]. Instead of manually creating a concept set, an LLM is used to generate concepts per disease. Then concepts are filtered via submodular optimization (which is a mathematical approach applied to a set of concepts and acts similar to feature selection techniques) and concepts with the highest visual activation scores are selected to train the final bottleneck layer.

### 2.3.2 CBMs for Vision-Language Models

For vision-language models, there's an emerging trend in generating initial textual concept sets using LLMs by prompting it for feature relevant concepts for each class, removing the need for manual concept set creation. Language guided bottlenecks (LaBo), leverages GPT-3 [17] to find factual concepts to first generate a pool of candidate concepts which are used to make even CLIP models inherently interpretable [134]. As the bottleneck layer restricts the model's internal parameters to only focus on understandable concepts, it allows the user to observe which concepts guide the final output generation. This idea is further leveraged to create a completely label-free CBM, shown in figure 2.7, where a Concept Bottleneck Layer is [91] learned via calculating a concept activation matrix and optimizing the projections weights to maximize the CLIP's performance. CLIP-dissect [92] is used to measure the similarity between activation patterns and target concepts.



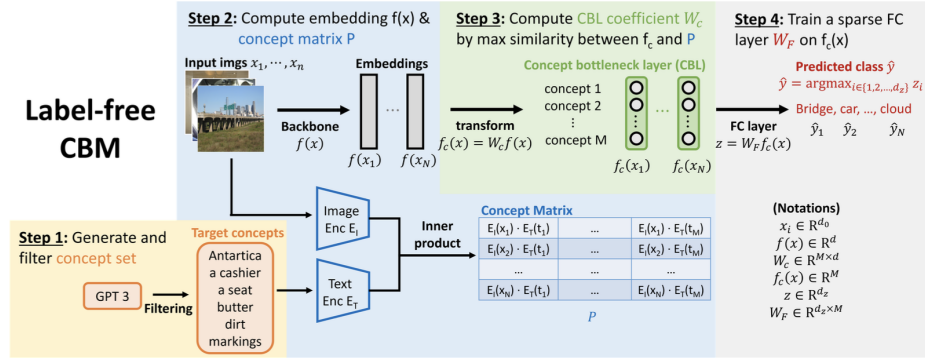


Figure 2.7: Label-free Concept Bottleneck Models [91] automate the process of text concept creation using GPT models. A CBL layer is then learned with the objective of maximizing similarity between the concept activation vector and the concept matrix, which is the joint representation of projection of image and concept embeddings. This helps the model focus on the most meaningful concepts by maximizing the alignment between the activations and the image-derived concepts. The final layer then minimizes the classification error with the objective to improve accuracy while maintaining interpretability through a sparse representation.

All of these works enforce sparsity in the concept bottleneck layer as it has been proven to be useful to enhance model’s interpretability. The same technique has also been used in many other works [94] accompanied by both contrastive and special loss functions such as Gumbel-softmax distribution based loss [111]. The notion of hierarchical concepts has also been reiterated in CBMs [16] to accommodate different granularities within the concept space provided in figure 2.8, which however comes at the expense of computational complexity and still requires broader evaluation across different tasks and datasets.

## 2.4 Domain-specific Approaches: Medical AI

In the context of application in medical imaging, classic methods like LRP, GRAD-CAM [110], SHAP, saliency maps with sanity checks for attribution methods are used for explaining deep learning methods for ECG analysis, heart-based diagnosis, lung disease classification, X-ray diagnosis and others [125, 40, 108, 133]. Self-explainable Neural Networks proposed [61] for skin lesion classification is an approach that fine-tunes CNNs and saliency map based explanations at the same time without affecting model’s performance.

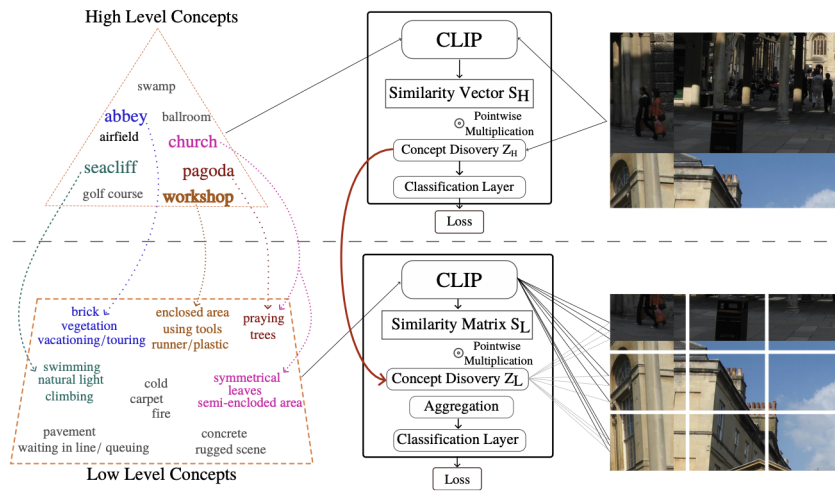


Figure 2.8: Concept Hierarchy [16] defined in two levels of granularity and compared to image patches using CLIP's contrastive learning approach.

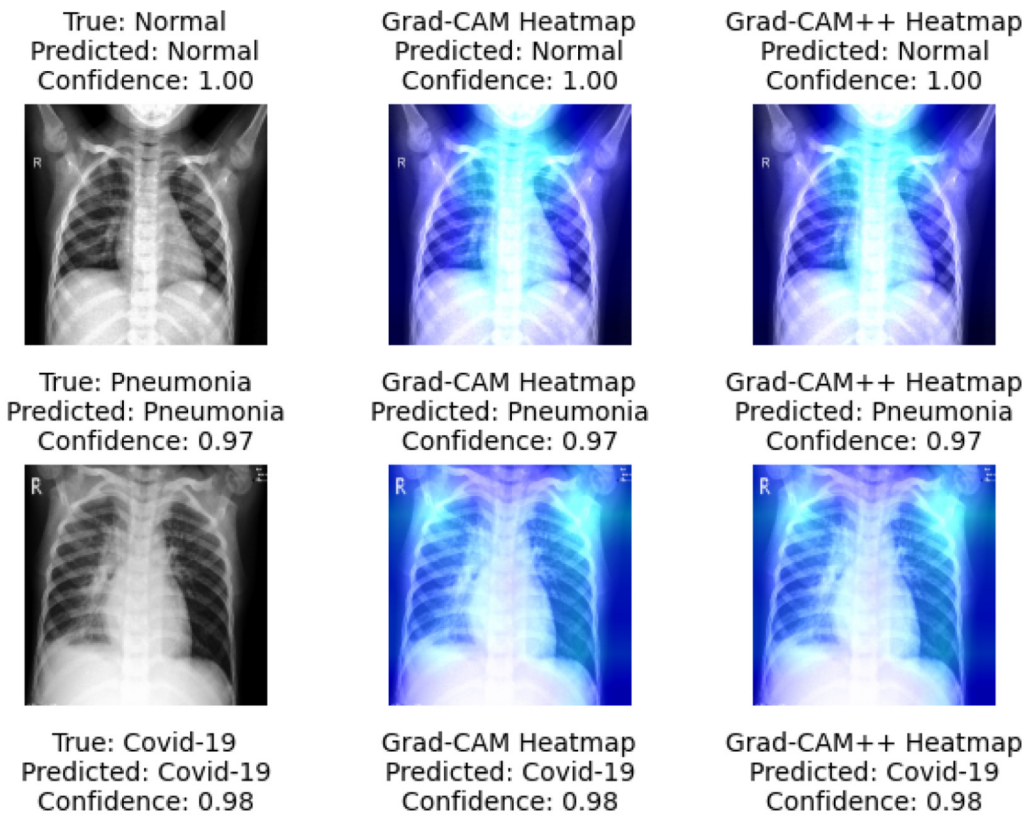


Figure 2.9: Grad-CAM based explanations for Medical Chest X-rays [87]

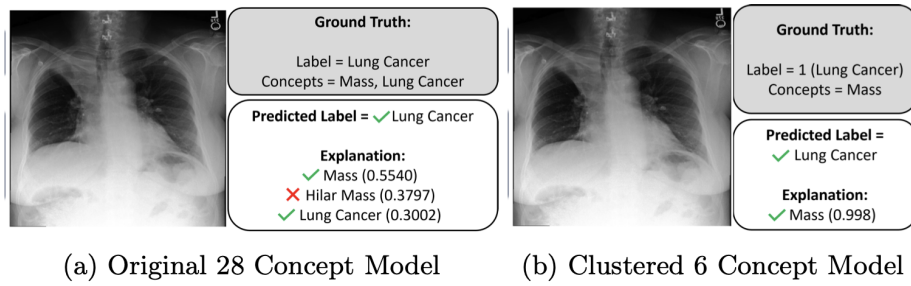


Figure 2.10: Concept based explanations for Medical Chest X-rays [100]  
train concept extractors for interpretable cancer classification

However, methods based on heatmaps do not provide clearly interpretable or distinguishable answers as to why the model made a certain decision [87]. Independent approaches have also been applied to explain medical decisions and support medical professionals. Xplainer [96] focuses on explaining zero-shot X-ray diagnosis by leveraging CLIP embeddings for both image and text reports, calculating their cosine similarities and providing an explanation based on the joint probability distribution of classes. Needless to say, the assumption of joint probability introduces a bias in this model which is not aligned with clinical assumptions. Recent work on Clinically-relevant Concept bottlenecks [19] show the application of CBMs in enhancing interpretability within deep learning models for lesion detection. Concept extraction for X-rays [100] for binary classification models have been shown to perform well against traditional XAI methods such as LIME or SHAP; example shown in figure 2.10. Model distillation techniques have been also applied [46] to find Post-hoc CBM models that can provide more meaningful explanations for medical x-ray data without hurting the performance of original black-box models. We found that these works and their practical implementation are limited to deep learning models [90] and has not been transferred to Vision-language models due to their complex nature. We also found research works on design principles that should be incorporated in XAI based medical decision support systems [18], which has not been implemented in any practical work so far despite having 77% of the medical practitioners willing to adopt such guidelines.

## 2.5 Summary

In general, we see although advanced methods exist to provide useful explanations, XAI has many dimensions (user-based, task-based) and therefore, is difficult to be transferred despite models having same architectural components. The human understandability factor is quite broad and needs to be accurately implemented to ensure usability of such explanations. Heatmaps, though widely used, are often not sufficiently interpretable for clinicians to fully trust or understand the model's decision-making process. There is a clear need for an understandable mapping between visual and textual correlations within vision-language models, as how they map multimodal representations to outputs is still unclear. Therefore, this research work aims to address these limitations by developing a methodology that automatically maps visual elements to text outputs in a way that is easily understandable to humans. The goal is to create an automated and easy-to-employ interpretability method, improving both usability and trust in AI-driven medical

diagnostics. To evaluate the approach, we also explore systematic evaluation frameworks to ensure consistency and trustworthiness. Earlier research works introduced various metrics to establish and verify the usefulness of explanations [60], while primary method remains to be Faithfulness. Faithfulness scores are introduced [118, 97] which calculates the correlation between explanations and the actual decision-making process of the model being explained. Understandability is another metric, seen in many works that relies on human evaluations and user feedback. Discriminability Scores are introduced to measure how well a concept aligns with images of a specific class, contributing to the concept selection process [66]. There's a high trade-off between Efficiency and Interpretability in CBMs as training the model requires higher computational cost. Furthermore, they often rely on a static initial set of concepts that is costly to generate for large datasets. Therefore, we also will evaluate the Efficiency [60] of the approach to understand how much model's response time is affected due to interpretable layer and if the trade-off is worth it. In this work, we aim to develop a methodology to -

- Automate Visual Concept Creation for Medical X-ray data
- Enhance Textual Concept Generation for Medical Reports
- Improve Concept Bottleneck Learning for Vision-Language Models (VLMs)
- Ensure Global Interpretability for Medical VLMs

---

# Chapter 3

## Technical Background

### 3.1 Transformers

Vaswani et al. [123] introduced the Transformer architecture in their seminal work *Attention Is All You Need*, which revolutionized deep learning by eliminating recurrence in favor of a self-attention mechanism. The Transformer processes sequences of embeddings in parallel and learns dependencies by weighing interactions between all token pairs through multi-head attention. This mechanism enables efficient modeling of long-range dependencies and facilitates parallel computation, resulting in faster training and scalability to large datasets. Their model outperformed previous methods on translation tasks such as English-to-German and English-to-French while requiring significantly less training time. Today, the Transformer serves as the backbone of modern AI systems, especially large language models.

The Transformer architecture consists of an encoder-decoder structure built upon layers of multi-head self-attention and position-wise feed-forward networks. The encoder transforms an input sequence into a set of continuous representations, while the decoder generates output sequences by attending to both encoder outputs and previous decoder outputs. A key innovation is the *multi-head self-attention* mechanism, which computes scaled dot-product attention in parallel across multiple heads. Each head allows the model to focus on different positions and semantic subspaces of the input. Stacking multiple layers of attention and feed-forward modules facilitates deep representations and efficient sequence modeling. Positional encodings are added to preserve order information, allowing full sequence-level parallelism during training and inference.

Transformers underpin a wide array of Natural Language Processing (NLP) systems including BERT [37], GPT [99], and other large language models [17]. These models achieve state-of-the-art performance in diverse tasks such as question answering, summarization, and machine translation. The parallelizable architecture enables scalable training on massive corpora, which has been instrumental in recent advancements in deep learning. In summary, the Transformer architecture introduced by Vaswani et al. [123] has become a foundational paradigm in contemporary AI research and deployment.

## 3.2 Vision Transformers (ViTs)

In 2021, Dosovitskiy et al. [42] introduced the Vision Transformer (ViT), demonstrating that pure Transformer architectures can achieve strong performance in image classification tasks without convolutional neural networks (CNNs). This challenged the long-standing dominance of CNNs, which build in strong inductive biases like spatial locality and translation equivariance [70]. In ViT, an image is divided into fixed-size patches (e.g.,  $16 \times 16$  pixels), each flattened and linearly projected into a latent vector. These patch embeddings are then treated as input tokens to the Transformer encoder. Self-attention is applied to these token sequences, capturing relationships across the entire image. ViTs establish that Transformers can be competitive or superior to CNNs, especially when trained on large-scale datasets due to their data-hungry nature.

The Vision Transformer mirrors the original Transformer encoder design. Each image patch is linearly embedded, and positional embeddings are added to retain spatial structure. A learnable [CLS] token is prepended to the sequence, and its final representation is used for classification. Layers of multi-head self-attention and feed-forward networks are applied in sequence. Unlike CNNs, ViTs lack built-in inductive biases, making them more reliant on extensive pre-training. While this allows ViTs to learn from data without pre-defined heuristics, they may underperform CNNs in low-data regimes. Nevertheless, ViTs have opened new avenues for fully attention-based modeling in vision.

## 3.3 Vision-Language Models (VLMs)

Vision-Language Models (VLMs) are designed to learn joint representations of visual and textual information, enabling a wide range of multimodal tasks such as image captioning, visual question answering (VQA), and automatic medical report generation. The ability to reason across modalities reflects the multimodal nature of real-world problems, particularly in domains like healthcare, where decision-making often requires both visual (e.g., X-rays, MRIs) and textual (e.g., clinical notes, diagnostic reports) contexts [48]. By processing inputs from both modalities, VLMs aim to develop more robust and contextually informed representations that improve task performance. This capability has opened the door to applications such as zero-shot image classification, multimodal retrieval, visual-question answering and automated diagnostic assistance [48].

### Architectural Design

The underlying architecture of VLMs can be broadly classified based on how and when cross-modal interactions are performed. The two dominant paradigms are *single-stream* and *dual-stream* architectures [24].

#### Single- vs. Dual-Stream Architectures

**Single-stream models**, such as VisualBERT [78] and UNITER [26], perform early fusion by concatenating visual embeddings and text tokens, feeding them into a unified transformer model. This setup enables the model to learn joint representations across modalities from the earliest stages of processing. Because the same parameter set is used for both image and text features, single-stream architectures are typically more parameter-efficient and computationally lightweight during training and inference.

In contrast, **dual-stream models**, including ViLBERT [84] and CLIP [98], adopt a late fusion strategy. Each modality is encoded independently through separate transformer-based encoders, and cross-modal interactions are introduced at later stages via attention mechanisms. This design preserves modality-specific information before integration, which can enhance performance on tasks that require fine-grained understanding of each modality. However, dual-stream models are usually more resource-intensive and complex to train [48].

### Encoder vs. Encoder-Decoder Frameworks

In addition to fusion strategies, VLMs can be differentiated by their use of encoder-only or encoder-decoder architectures [24].

**Encoder-only models**, such as ALIGN [55], are typically used for retrieval tasks, where the objective is to map visual and textual inputs into a shared embedding space. These models excel in efficiency and scalability, making them well-suited for real-time applications like multimodal search and contrastive learning. However, their utility in generative scenarios is limited due to the absence of a decoding mechanism.

On the other hand, **encoder-decoder models**, exemplified by SimVLM [127], are designed to support free-form text generation from multimodal inputs. The encoder aggregates features from both modalities into a unified representation, which the decoder then uses to generate outputs such as captions or long-form narratives. Although encoder-decoder models offer greater expressive capacity for generation, they also impose increased computational demands and longer inference times.

Overall, the choice of architecture—whether single-stream or dual-stream, encoder-only or encoder-decoder—reflects a trade-off between performance, efficiency, and applicability across different multimodal tasks. These architectural choices continue to shape the research and development of increasingly capable vision-language models.

### Training Paradigms for VLMs

Training VLMs effectively requires strategies that allow the model to learn robust and generalizable cross-modal representations. Several paradigms have been adopted:

- **Transfer Learning:** Pre-training on large-scale datasets followed by fine-tuning on task-specific data is a common approach. Models such as CLIP [98] demonstrate the effectiveness of learning from noisy but extensive image-text pairs.
- **Curriculum Learning:** Some medical VLMs (e.g., LLaVa-Med [74]) employ curriculum learning strategies, presenting simpler tasks early in training and gradually increasing complexity to improve model generalization.
- **Self-Supervised Learning (SSL):** SSL enables models to learn without relying on large amounts of labeled data, which is particularly valuable in medical contexts where annotations are costly. Common pretext tasks include contrastive learning [122] and masked modeling [37, 131].
- **Multitask Pretraining:** Many modern VLMs combine multiple training objectives (e.g., image-text matching, masked language modeling) to enrich the learned representations and improve performance across downstream tasks [77].

## Challenges and Advancements in Medical VLMs

Adapting general-domain Vision-Language Models (VLMs) to clinical applications presents unique challenges due to the specialized, domain-specific, and sensitive nature of medical data [48]. Ensuring clinically accurate and trustworthy outputs necessitates robust interpretability mechanisms and fine-grained visual-textual understanding.

To address these challenges, recent advancements have been made in the development of medical VLMs. One such advancement is the introduction of LVM-Med, a large-scale self-supervised vision model tailored for medical imaging. LVM-Med leverages a novel graph-matching formulation to enhance representation learning across diverse medical imaging modalities, demonstrating superior performance on various downstream tasks [89].

Additionally, advancements have been made in biomedical claim verification through the integration of large language models (LLMs), transparent model explanations, and user-guided justification. An interactive biomedical claim verification system has been developed to classify scientific studies as "Support," "Contradict," or "Not Enough Information" regarding specific claims, enhancing the transparency and interpretability of AI-assisted decision-making in biomedical contexts [81].

Overall, vision-language modeling represents a critical area for advancing multimodal artificial intelligence, particularly in complex, high-stakes fields such as healthcare. Continued innovation in model architectures, training strategies, and evaluation methods is essential to fully realize their potential in supporting clinical decision-making and improving patient care.

## 3.4 XAI for Vision-Language Models

Vision-Language Models (VLMs) integrate two distinct data modalities—visual embeddings (e.g., from CNNs or Vision Transformers) and text embeddings (e.g., from transformers or language models)—via shared latent spaces or cross-modal fusion. The complexity of these architectures poses a challenge for transparency. To address this, Explainable Artificial Intelligence (XAI) techniques have been adapted to for these pipelines, including attention-based saliency, attribution methods, and intermediate concept supervision. In this work, we focus on concept bottleneck models (CBM), which are one of the inherent interpretability paradigms. Below we provide the theoretical foundations of a CBM:

### 3.4.1 Concept Bottleneck Models for VLMs

Concept Bottleneck Models (CBMs) introduce an interpretable layer between the representation and decision layers by explicitly using human-understandable concepts before classification [69]. In the standard CBM framework, the model is structured as a composition of two functions: a concept encoder  $g : \mathcal{X} \rightarrow \mathbb{R}^K$ , which maps the input  $x \in \mathcal{X}$  to a concept vector  $\mathbf{c} \in \mathbb{R}^K$ , and a downstream classifier  $f : \mathbb{R}^K \rightarrow \mathbb{R}^C$ , which maps concepts to class probabilities.

#### Training Objective:

The overall model is defined as:

$$\hat{y} = f(g(x))$$



The training objective minimizes the sum of two loss terms:

$$\mathcal{L}_{\text{CBM}} = \lambda_1 \cdot \mathcal{L}_{\text{concept}}(g(x), c) + \lambda_2 \cdot \mathcal{L}_{\text{task}}(f(g(x)), y)$$

where  $\mathcal{L}_{\text{concept}}$  is the loss between predicted concepts  $g(x)$  and ground truth concept labels  $c \in \{0, 1\}^K$ ,  $\mathcal{L}_{\text{task}}$  is the task loss, typically cross-entropy between predicted class logits and ground truth labels  $y \in \{1, \dots, C\}$  and  $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$  are scalar weights to balance the two objectives.

In scenarios without concept annotations, the concept loss is replaced by an unsupervised or weakly-supervised objective (e.g., alignment-based or sparsity-promoting terms), forming a *label-free* CBM [91]. The intermediate concept vector  $c$  serves as an interpretable bottleneck. Each dimension  $c_k$  corresponds to a semantically meaningful concept (e.g., "a round face", "large teeth"). By inspecting  $c$  and the weights of  $f$ , one can trace predictions to high-level abstractions, enabling both global (model-level) and local (instance-level) explanations.

To enhance post-hoc transparency, natural language rationales can be generated alongside predictions [91]. Retrieval-Augmented Generation (RAG) architectures extend CBMs by introducing an evidence aggregation step. Visual concepts are first mapped to textual findings, which are passed to a generative module conditioned on external knowledge sources. Recent works combine CBM outputs with RAG-style decoders to produce justification chains for medical classifications [5]. This augments structural interpretability with narrative-level explanations, aligning the system’s reasoning with clinical diagnostic procedures. While attention and attribution methods remain foundational, they provide only surface-level insights into model focus and saliency. CBMs introduce architectural interpretability by enforcing concept-level supervision and representation disentanglement, which is particularly effective in domains with well-defined ontologies. Integrating CBMs with VLMs enables a unified framework that supports interpretable intermediate reasoning and multimodal explanation, and hence is chosen for this work.

### 3.5 Evaluation Metrics

Evaluating the quality of explanations produced by Vision-Language Models (VLMs) is crucial for ensuring their reliability, especially in high-stakes domains like medical diagnostics. Unlike traditional performance metrics (e.g., accuracy, F1-score), explainability metrics assess how well a model’s reasoning aligns with human understanding. This section outlines key evaluation metrics pertinent to explainable AI (XAI) in VLMs. Faithfulness measures the extent to which an explanation accurately reflects the model’s decision-making process. A faithful explanation should correspond directly to the internal computations of the model. Metrics used to assess faithfulness include:

- **Deletion and Insertion Metrics:** Evaluate the impact on model output when important features identified by the explanation are removed or inserted [106].
- **Sensitivity Analysis:** Measures how sensitive the model’s predictions are to changes in input features deemed important by the explanation [10].

In this work, we used deletion and insertion based Area Under the Curve (AUC) measurements to address faithfulness. Plausibility is another metric that assesses how convincing or understandable an explanation is to human users, regardless of its faithfulness. Common evaluation methods include human judgment studies where participants rate the

quality of explanations based on clarity and usefulness [41]. Agreement with human annotations compares model-generated explanations with human-annotated rationales [38]. Under the scope of this work, we do not conduct any human studies. However, the visual concepts are understandable by any human expert radiologists as they are directly derived from the radiographs. On the other hand, generating textual concepts as described in section 4.2 is a well-established approach adopted by many research works.

In Concept Bottleneck Models (CBMs), it's essential to evaluate the quality of the intermediate concepts. Therefore concept Completeness is measured to understand how well the set of concepts captures all the information necessary for the final prediction [136]. Concept Purity is also assessed to know whether each concept corresponds to a distinct, interpretable feature without overlap [45]. We use an interpretability heuristic to determine concept purity instead of adapting to any filtering techniques (see section 5.4 Consistency on the other hand examines whether the model provides similar explanations for similar inputs. We run the models multiple times and consider the mean and standard deviation to define stability Metrics [8].

In addition to quantitative metrics, human-centered evaluations are vital. User studies assess how explanations affect user trust, satisfaction, and decision-making whereas task performance measures whether explanations help users perform tasks more effectively. Although we acknowledge that in addition to quantitative metrics, human-centered evaluations are vital, it has not been carried out during the thesis phase.

## Challenges and Future Directions

Despite advancements, evaluating explanations in VLMs remains challenging due to lack of standardized benchmarks where diverse tasks and datasets make it hard to compare methods. Subjectivity in human evaluations can also vary, affecting the reliability of plausibility assessments. In this work, we provide qualitative explanations and analysis that may help clinicians evaluate the plausibility, faithfulness and reliability of proposed approaches.

## Chapter 4

# Methodology

In this work, we propose a two-phase framework for learning interpretable visual-language representations using Concept Bottleneck Models (CBMs). The proposed approach decomposes the end-to-end prediction pipeline into semantically structured stages that explicitly incorporate human-understandable visual and textual concepts. The goal is to create a model that not only performs well on medical classification tasks but also provides concept-level explanations in both modalities.

The methodology comprises two main components, as illustrated in Figure 4.1: (i) *Concept Set Creation* and (ii) *Learning the Concept Bottleneck Layer*. Each phase is designed to enforce interpretability constraints while retaining predictive performance.

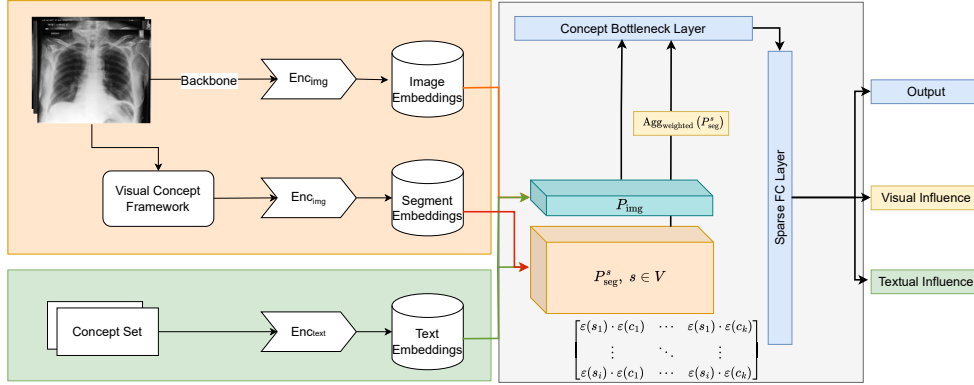


Figure 4.1: Overview of the proposed visual-language concept bottleneck framework. Visual concepts are extracted via a trained and interpretable concept generator module. A concept bottleneck layer, trained with a sparse classifier, maps these concepts to the final task output. The framework also supports influence analysis to quantify visual and textual contributions.

**Phase I: Concept Set Generation** begins with the extraction of high-dimensional visual embeddings from a pre-trained backbone network applied to input X-rays. A

concept generation module segments the image into semantically meaningful parts. Each segmented region is then embedded into a feature space using a visual encoder. Simultaneously, a textual concept set—representing domain-specific diagnostic terms—is embedded using a text encoder, creating a shared latent space for cross-modal comparison. The concept set is not pre-annotated but is generated and processed automatically using prompt engineering.

**Phase II: Learning the Bottleneck Layer** focuses on constructing the concept bottleneck layer. Visual segments and text concepts are projected into a shared embedding space. For each image, a *Concept Matrix* is constructed, where each entry captures the alignment score between a segmented visual region and a concept. These alignment scores are computed using a similarity function between their respective embeddings. An aggregation function, such as weighted pooling over visual segments, is applied to obtain a final concept activation vector for the image. This serves as the input to the bottleneck layer. The concept bottleneck vector is passed through a sparse fully connected classifier that maps interpretable concept activations to final diagnostic predictions.

The output of this module supports both quantitative analysis and human-readable explanations, enabling insight into which visual concepts were most influential for a given prediction. In the subsequent sections, we describe each component in detail, including the tailored loss objectives, training procedure, and evaluation metrics.

## 4.1 Visual Concept Generation

Automatic extraction of human-understandable visual concepts from raw X-ray data serves as the foundational step in our proposed approach, particularly in frameworks designed to elevate the interpretability of deep neural networks (DNNs). This section introduces the concept generation phase in our implementation, which is adopted from the Explain Any Concept (EAC) framework [118].

To overcome the limitations of manual annotation and fragile unsupervised clustering, we utilize the Segment Anything Model (SAM) [67], an instance segmentation framework, for automated and scalable concept discovery. This integration enables us to construct concept sets from input images with high fidelity and semantic richness. These regions are subsequently refined via a filtering pipeline to remove redundancies and artifacts, yielding a structured and interpretable concept set  $C = \{c_1, c_2, \dots, c_n\}$ .

Visual concept generation constitutes the first phase of a two-stage pipeline in our framework. The stages are as follows:

1. **Concept Discovery:** SAM segments an input image into multiple object-level regions, which are proposed as visual concepts.
2. **Explanation via Shapley Estimation:** These filtered concepts are passed to a surrogate model trained using the Per-Input Equivalence (PIE) [118] scheme, where Shapley values [85] are computed to quantify concept importance.

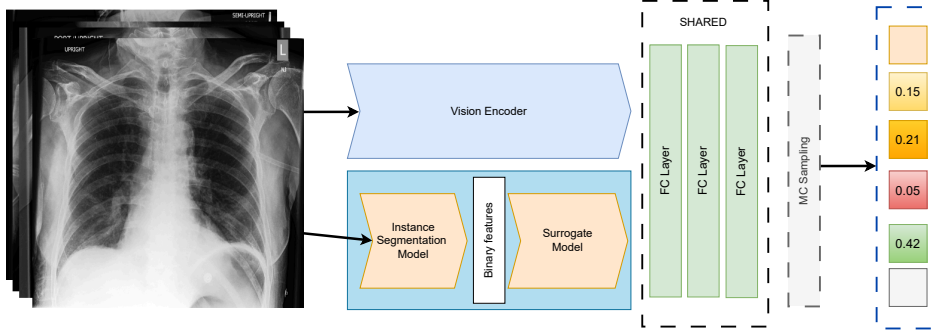


Figure 4.2: Overview of the proposed Visual concept generation framework.

While effective on natural images, the quality of concept extraction depends heavily on the domain relevance of the segmentation model. Therefore, for medical imaging tasks, we used MedSAM [86] which is a Segmentation model trained with a very large-scale medical image dataset, covering 10 imaging modalities.

We elaborate on the segmentation approach in the following section 4.1.1.

#### 4.1.1 Concept Discovery via Segmentation

To derive semantically meaningful visual concepts from raw images, we employ a segmentation-based approach rooted in pre-trained foundation models. Let an input image be denoted as  $x \in \mathbb{R}^{H \times W \times 3}$ . The segmentation model  $\mathcal{S}$  maps  $x$  to a finite set of binary masks:

$$\mathcal{S}(x) = \{m_1, m_2, \dots, m_N\}, \quad m_i \in \{0, 1\}^{H \times W}, \quad (4.1)$$

where each mask  $m_i$  defines a distinct spatial region in the image that is hypothesized to correspond to a semantically coherent object or object part.

We instantiate  $\mathcal{S}$  using either the Segment Anything Model (SAM) [67] or a domain-specific variant such as MedSAM, depending on the application domain. These models are applied in zero-shot inference mode without any task-specific fine-tuning.

The resulting collection  $\mathcal{C}(x) = \{c_1, \dots, c_N\}$ , where each  $c_i = m_i \odot x$  denotes the masked image region, is defined as the *visual concept set* for image  $x$ . Importantly, no post-hoc filtering is applied to this set. All masks  $\{m_i\}$  produced by  $\mathcal{S}$  are retained for subsequent analysis. This decision ensures maximal fidelity to the segmentation model’s predictions and avoids introducing inductive bias via handcrafted filtering criteria.

In implementation, all segments are cached to avoid recomputation across multiple explanation runs. The visual concept set  $\mathcal{C}(x)$  forms the basis for downstream perturbation and Shapley-based attribution analysis, as detailed in Sections 4.1.2 and 4.1.3. This fully automated segmentation-based procedure enables scalable and architecture-agnostic concept generation across diverse visual domains.

#### 4.1.2 Per-Input Equivalence (PIE)

To enable efficient computation of concept-level attributions, we adopted a surrogate modeling technique termed *Per-Input Equivalence* (PIE) [118]. Rather than relying on

the full target model  $f$  for evaluating the marginal contribution of each concept subset—which is computationally expensive—we construct a simplified surrogate model  $f'$  that is trained to approximate  $f$ 's behavior for a specific input image.

Let the concept set for an image  $x$  be denoted as  $\mathcal{C}(x) = \{c_1, \dots, c_N\}$ . We define a binary selection vector  $z \in \{0, 1\}^N$ , where  $z_i = 1$  indicates that concept  $c_i$  is preserved in the input and  $z_i = 0$  that it is masked out. For each  $z$ , we synthesize a masked image:

$$x_z = \sum_{i=1}^N z_i \cdot c_i, \quad (4.2)$$

and compute the feature embedding  $\phi(x_z) \in \mathbb{R}^d$  via a fixed visual encoder.

We then train a surrogate model  $f' : \{0, 1\}^N \rightarrow \mathbb{R}^K$  such that:

$$f'(z) \approx f(x_z), \quad (4.3)$$

where  $f(x_z)$  is the predicted probability distribution from the original model, and  $K$  is the number of classes. The surrogate  $f'$  is instantiated as a shallow neural network that learns to map binary vectors  $z$  to output logits, using cross-entropy loss over 2500 Monte Carlo-sampled combinations of concepts.

To further constrain  $f'$  and enhance approximation fidelity, the final fully connected layer of the target model  $f$  is reused (parameter sharing). This allows the surrogate to preserve the original class decision boundaries while decoupling it from the costly visual backbone.

Formally, the surrogate model is defined as:

$$f'(z) = \text{FC}(\psi(z)), \quad (4.4)$$

where  $\psi$  is a learnable nonlinear transformation and FC is the frozen final linear layer of  $f$ .

Training proceeds by minimizing the KL divergence between the class probability vectors  $f(x_z)$  and  $f'(z)$  across a dataset of sampled binary vectors:

$$\min_{\psi} \mathbb{E}_{z \sim \text{Bernoulli}(p)} [D_{\text{KL}}(f(x_z) || f'(z))]. \quad (4.5)$$

Once trained, the surrogate  $f'$  is used in place of  $f$  to estimate the Shapley value of each concept efficiently. This substitution reduces computational cost by orders of magnitude while preserving input-specific behavioural equivalence.

### 4.1.3 Shapley-Based Attribution

To quantify the influence of each visual concept on the model's prediction, we adopt a Monte Carlo approximation of the Shapley value [85], a principled measure from cooperative game theory. Let  $\mathcal{C}(x) = \{c_1, \dots, c_N\}$  be the set of visual concepts derived for input  $x$ , and let  $f'$  be the per-input surrogate model trained as described in Section 4.1.2. For a given class of interest  $y$ , we seek to estimate the marginal contribution of each concept  $c_i$  to the class probability  $f'(z)_y$ .

Following the definition of the Shapley value, the contribution of concept  $c_i$  is computed as:

$$\phi_i = \mathbb{E}_{S \subseteq \mathcal{C} \setminus \{c_i\}} [f'(S \cup \{c_i\})_y - f'(S)_y], \quad (4.6)$$

where  $S$  denotes a randomly sampled subset of the concept set excluding  $c_i$ . Since the number of such subsets is exponential in  $N$ , we use Monte Carlo sampling to approximate the expectation.

For each concept  $c_i$ , we independently sample  $K$  binary vectors  $z^{(1)}, \dots, z^{(K)} \in \{0, 1\}^N$  such that:

$$z_j^{(k)} \sim \text{Bernoulli}(0.5), \quad \text{for } j \neq i. \quad (4.7)$$

Two variants of each sample are created:

$$z_{+i}^{(k)} = z^{(k)} \text{ with } z_i^{(k)} = 1, \quad (4.8)$$

$$z_{-i}^{(k)} = z^{(k)} \text{ with } z_i^{(k)} = 0. \quad (4.9)$$

The Shapley value for  $c_i$  is then estimated as:

$$\hat{\phi}_i = \frac{1}{K} \sum_{k=1}^K \left[ f'(z_{+i}^{(k)})_y - f'(z_{-i}^{(k)})_y \right]. \quad (4.10)$$

In practice, we set  $K = 50,000$  to ensure convergence of the Monte Carlo estimator. This sampling procedure is repeated independently for each concept, resulting in a vector of Shapley values  $\hat{\phi} \in \mathbb{R}^N$ .

The resulting scores are sorted to identify the most influential concepts. The top- $k$  concepts are then saved as binary masks for later use. This process enables local interpretability and visual transparency by grounding explanations in identifiable image regions.

## 4.2 Textual Concept Extraction via Prompting

Textual concepts are extracted using a structured prompt applied to the OpenAI GPT [1] model via API. The prompt is designed to elicit short, radiologically relevant phrases grounded in visible findings observable in X-ray images. The prompt follows the form:

```
You are a radiology assistant extracting concise visual concepts
from X-rays for diagnostic purposes. [...] Format the output
strictly as a Python list.
```

The full prompt can be found in the appendix.

Each diagnostic label  $l_i$  from a dataset is passed as input to the prompt template. The LLM was instructed to avoid using non-visual concepts. The response from the language model is parsed and stored as a Python list of string concepts. These lists are reused for downstream concept bottleneck supervision and for evaluation of visual-to-text alignment.

Formally, given a label  $l_i$ , the LLM produces:

$$T_i = \text{LLM}(\text{prompt}(l_i)), \quad T_i = [t_1, t_2, \dots, t_n], \quad (4.11)$$

where  $T_i$  is a list of textual concepts describing radiologically relevant visual patterns associated with  $l_i$ .

## 4.3 Concept Bottleneck Modeling (CBM)

### 4.3.1 Architecture Overview

Our Concept Bottleneck Model (CBM) is designed to learn a semantically structured representation of images by projecting high-dimensional vision-language embeddings into a lower-dimensional bottleneck space composed of disentangled visual and textual concepts. The architecture adheres to the encoder-bottleneck-decoder paradigm, with the key novelty lying in the design of the bottleneck supervision and loss composition:

**Encoder.** Given an input image  $x \in \mathbb{R}^{H \times W \times 3}$ , a pre-trained CLIP-based vision encoder  $f_{\text{clip}}$  maps it to a latent feature vector  $z = f_{\text{clip}}(x) \in \mathbb{R}^d$ . These latent vectors are precomputed and cached prior to CBM training.

**Projection Layer.** The feature vector  $z$  is passed through a projection module  $\mathcal{P}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^C$ , where  $C$  is the number of learned concept dimensions. This layer is a learnable fully connected network optimized via supervised alignment losses.

**Concept Bottleneck Supervision.** Each projected concept embedding is supervised using both visual concepts (see Section 4.1) and textual concepts (see Section 4.2). The bottleneck is structured to encourage alignment between corresponding visual and textual representations of concepts.

**Training Objective.** To learn the projection layer, we define a composite loss function incorporating alignment with textual and visual concepts. These loss components are introduced in detail in Section 4.3.2, and briefly summarized here:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{icl}} + \lambda \mathcal{L}_{\text{agg}} + \mathcal{L}_{\text{out}}, \quad (4.12)$$

where  $\mathcal{L}_{\text{icl}}$  is a local concept alignment loss,  $\mathcal{L}_{\text{agg}}$  is an aggregated feature similarity loss, and  $\mathcal{L}_{\text{out}}$  is an output similarity loss. Each term is weighted using user-specified hyperparameters  $\alpha$  and  $\lambda$ .

The CBM is trained using mini-batch gradient descent over randomly sampled subsets of concept embeddings. The original image encoder is frozen, and only the projection layer parameters  $\theta$  are updated. Training proceeds for a fixed number of projection steps using a pre-defined batch size and learning rate. Once trained, the projection layer serves as input to a sparse final classifier detailed in Section 4.3.2.

### 4.3.2 Visual-Textual Concept Supervision

The learning phase of the Concept Bottleneck Model (CBM) relies on semantically aligned supervision from both visual and textual modalities. As illustrated in Fig. 4.1, our approach uses segmentation-derived visual concept embeddings and prompt-generated textual concept descriptors to enforce structured learning in the bottleneck layer.



### Visual Concept Embeddings

From each input image  $X$ , the *Concept Generation Module* extracts a set of  $V$  visual concept masks via the segmentation model. Each mask  $v_i \in \{0, 1\}^{H \times W}$  is applied to the image to isolate a region, which is then passed through the vision module of the pretrained VLM to yield a visual embedding:

$$F_{\text{VM}}(v_i; W_{\text{VLM}}), \quad v_i \in \mathcal{V}_X, \quad (4.13)$$

where  $\mathcal{V}_X$  is the set of all visual concept masks for  $X$ . These embeddings are stacked to form the **visual concept matrix**  $V = F_{\text{VM}}(\mathcal{V}_X; W_{\text{VLM}})$ .

### Textual Concept Embeddings

For each diagnostic label, a list of textual phrases is generated using a structured prompt applied to an LLM (see Section 4.2). Each phrase  $c_j$  is embedded using the text encoder of the pretrained VLM:

$$F_{\text{LM}}(c_j; W_{\text{VLM}}), \quad c_j \in \mathcal{C}, \quad (4.14)$$

forming the textual concept matrix  $C = F_{\text{LM}}(\mathcal{C}; W_{\text{VLM}})$ .

### Concept Bottleneck Layer

The Concept Bottleneck Model (CBM) includes two stages: a projection layer that computes similarity-based concept activations, and a sparse linear classifier that maps these activations to output labels.

**Projection Layer.** Let  $Z \in \mathbb{R}^D$  denote the global image embedding obtained from a pretrained vision-language model (VLM). We have already defined  $V = \{v_i\}_{i=1}^n \subset \mathbb{R}^D$  as the set of visual concept embeddings and  $C = \{c_j\}_{j=1}^m \subset \mathbb{R}^D$  represent the textual concept embeddings.

The projection layer computes pairwise similarities between visual/textual modalities using inner products:

$$P_{\text{img}} = ZC^\top, \quad (4.15)$$

$$P_{\text{seg}} = VC^\top, \quad (4.16)$$

where  $P_{\text{img}} \in \mathbb{R}^m$  is the Projection Matrix of full image embeddings and  $P_{\text{seg}} \in \mathbb{R}^{n \times m}$  represent the image-to-text and segment-to-text similarity matrices, respectively. The resulting concept activations are defined by flattening and/or aggregating these matrices.

$$f_{\text{CBL}}(Z, P_{\text{img}}, P_{\text{seg}}; W_c) = W_c h_\theta(Z, P_{\text{img}}, P_{\text{seg}}), \quad (4.17)$$

**Sparse Linear Classifier.** Given the concept activation vector  $h_\theta(Z, P_{\text{img}}, P_{\text{seg}}) = h_\theta(Z_c)$ , a sparse linear classifier produces the final prediction:

$$f_{\text{CBM}}(Z_c; W_g) = W_g h_\theta(Z_c), \quad (4.18)$$

where  $W_g \in \mathbb{R}^{K \times (n \cdot m)}$  is the weight matrix mapping concept activations to  $K$  output classes. Following projection layer training,  $W_g$  is optimized independently using the SAGA [35] optimizer with elastic-net regularization. The use of  $\ell_1$  and  $\ell_2$  penalties encourages sparse yet stable weights, supporting modular and interpretable decision-making.

## Loss Composition

Training is supervised by three complementary loss terms:

**(i) Output Similarity Loss ( $\mathcal{L}_{\text{out}}$ ).** To encourage the bottleneck projection to resemble the original CLIP space, the projected vector is compared against the original image embedding using a cosine similarity cube:

$$\mathcal{L}_{\text{out}} = -\cos^3(h_\theta(z), z). \quad (4.19)$$

**(ii) Local Concept Alignment Loss ( $\mathcal{L}_{\text{lcl}}$ ).** A similarity matrix is computed between each visual segment  $s_i$  and textual concept  $c_j$ , and a softmax weighting is used to compute a contrastive alignment loss:

$$\mathcal{L}_{\text{lcl}} = -\frac{1}{V} \sum_{i=1}^V \sum_{j=1}^C \text{softmax}(\varepsilon(s_i) \cdot \varepsilon(c_j)) \cdot (\varepsilon(s_i) \cdot \varepsilon(c_j)). \quad (4.20)$$

**(iii) Aggregated Feature Similarity Loss ( $\mathcal{L}_{\text{agg}}$ ).** To encourage the projected embeddings to be similar to the most representative visual segment, we compute an aggregated concept vector using max-pooling over  $P^{\text{seg}}$ :

$$\mathcal{L}_{\text{agg}} = -\cos^3(\text{Agg}_{\text{max}}(P^{\text{seg}}), h_\theta(z)). \quad (4.21)$$

## Final Training Objective

The total training loss is a weighted sum of the above three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{out}} + \lambda_{\text{lcl}} \mathcal{L}_{\text{lcl}} + \lambda_{\text{agg}} \mathcal{L}_{\text{agg}}. \quad (4.22)$$

The hyperparameters  $\lambda_{\text{lcl}}$  and  $\lambda_{\text{agg}}$  control the influence of concept alignment and feature similarity during optimization. All image encoders are frozen during training, and only the bottleneck projection layer  $W_{\text{CBL}}$  is updated using stochastic gradient descent over mini-batches.

In summary, the CBM training pipeline proceeds as follows: (i) visual and textual concepts are extracted and embedded; (ii) image-level CLIP embeddings are projected into the concept space via a learnable bottleneck; (iii) this projection is supervised using a combination of alignment and similarity losses; and (iv) the resulting normalized concept activations are used to train a sparse linear classifier for final prediction. This end-to-end supervision enables concept disentanglement while preserving classification performance.

## Understanding Sparsity

To quantify the interpretability of the final classifier, we measure the sparsity of  $W_g$  by computing the proportion of non-zero weights:

$$\text{sparsity}(W_g) = \frac{\|\text{nonzero}(W_g)\|_0}{\|W_g\|_0} \quad (4.23)$$

This allows us to monitor how many concepts are actually used by the final model, allowing for interpretability analysis and concept-level importance attribution. This final sparse layer completes the concept bottleneck pipeline, allowing the model to make predictions based on a compact, interpretable set of visual and textual concepts. Unlike conventional deep networks, the learned weights in  $W_g$  can be directly examined to understand which concepts contribute most to the model’s output.

## 4.4 Design Considerations

A critical design objective of in this framework is to balance interpretability and fidelity with computational efficiency. One of the strengths of our implementation is the use of pre-trained SAM models in a purely inference setting. No fine-tuning, retraining, or domain-specific calibration is required. This design decision aligns with the objective of creating a scalable and generalizable explainability system that can be readily applied to new datasets without extensive engineering overhead. The segmentation step is intentionally kept simple and modular. All detected instance masks are retained in their raw form without any filtering or merging. This maximizes compatibility with the subsequent explanation stage and ensures the reproducibility of concept attribution scores derived later via the Shapley value approximation.

Generating concept-level attributions via exact Shapley value computation over the original deep neural network  $f$  would be highly expensive due to two factors: (i) the exponential number of concept subsets, and (ii) the high cost of forward passes through  $f$  for each masked input.

To address this, these approach integrates three key efficiency mechanisms:

**1. Surrogate Approximation via PIE.** As described in Section 4.1.2, we used a per-input surrogate model  $f'$  trained to approximate  $f$  on concept-masked inputs. Because  $f'$  replaces the full forward pass through the visual encoder with a lightweight transformation of binary vectors, it substantially reduces the runtime of Shapley estimation. Let  $T_f$  and  $T_{f'}$  denote the average evaluation times of  $f$  and  $f'$  respectively; then the computational gain per sample is approximately:

$$\eta = \frac{T_f}{T_{f'}} \gg 1. \quad (4.24)$$

In empirical runs,  $T_f$  exceeds 2 seconds per input (with visual backbone and classification head), whereas  $f'$  completes inference in less than 10 milliseconds, resulting in a speedup of over  $200\times$  during MC sampling.

**2. Monte Carlo Sampling.** Instead of enumerating all  $2^N$  concept subsets, we estimate each Shapley value using  $K = 50,000$  random samples per concept. The use of binary sampling vectors enables efficient batching during inference, leveraging matrix-multiplication-optimized GPU execution.

**3. Result Caching.** To further accelerate the workflow and support reproducibility, all segmentation masks and surrogate models are cached on disk. Let  $\mathcal{M}$  be the segmentation cache mapping image paths to concept sets. Once a concept set  $\mathcal{C}(x)$  is computed for input  $x$ , it is stored in  $\mathcal{M}$  and reused across all subsequent explanation runs. This

---

prevents redundant invocations of the SAM or MedSAM models, whose runtime per image ranges between 1–3 seconds depending on resolution and hardware.

**4. Avoiding Redundant Forward Passes.** During the training of the surrogate  $f'$ , we precompute the masked features using the full visual encoder and store them. This avoids repeated feature extraction for each concept subset, amortizing the cost of the encoder over the sampled dataset.

Collectively, these optimizations make it feasible to generate concept-level Shapley explanations for high-resolution images in under one minute per instance. This contrasts with the naive baseline—which would require several hours per image due to repeated evaluation of a deep model on thousands of perturbed inputs. Our approach thus achieves a favorable trade-off between computational efficiency and explanatory fidelity.

---

# Chapter 5

## Experiments and Evaluation

### 5.1 Dataset Description

To evaluate our explainable framework across both general and domain-specific settings, we use five publicly available datasets that span natural image classification and medical imaging domains. These datasets are described below:

**ImageNet [36]:** ImageNet is a large-scale image classification dataset containing over 1.2 million images across 1,000 object categories. It serves as a benchmark for evaluating high-level vision models and is used here to assess generalizability to complex object-centered scenes.

**COCO [83]:** The Microsoft Common Objects in Context (COCO) dataset consists of 330k images with over 80 object classes, annotated for detection, segmentation, and captioning tasks. We use the validation set for concept-level attribution on richly annotated scenes.

**CheXpert [21]:** CheXpert Plus is a medical imaging dataset released by the Stanford AIMI Center, containing 223,228 unique chest X-ray images; both frontal-view and lateral view X-rays are labeled with 14 chest pathological conditions. We use the CheXbert labels to curate a single labeled dataset to evaluate the classification performance of our approach.

**MIMIC-CXR [56]:** MIMIC-CXR is derived from the MIMIC-IV clinical database and includes over 377,110 chest radiographs and corresponding reports. The dataset also provides 14 chest diseases from real-world hospital imaging data. We used the LongTail-CXR [50] dataset which provides single labels for each radiograph combining MIMIC-CXR and ChestX-ray8 [126] datasets, to select single-labelled radiographs.

Table 5.1: Total sample counts for each finding across MIMIC-CXR, CheXpert, and ChestX-ray8 datasets

Finding	MIMIC-CXR	CheXpert	ChestX-ray8
Total	250,022	191,229	108,948
No finding / Normal	83,271	17,000	84,312
Enlarged cardiomeastinum	7,915	9,132	-
Cardiomegaly	49,249	23,451	1,010
Airspace opacity	56,017	94,328	-
Lung lesion	7,058	6,997	-
Edema	29,560	49,717	-
Consolidation	11,813	13,015	-
Pneumonia	18,434	4,683	1,062
Atelectasis	49,960	58,777	5,780
Pneumothorax	11,674	17,700	2,793
Pleural effusion	59,104	76,963	-
Pleural (other)	2,166	2,506	-
Fracture	5,044	7,436	-
Support devices	74,698	107,269	-
Infiltration	-	-	10,317
Mass	-	-	6,046
Nodule	-	-	1,971

**COVID-QU-Ex [29]:** COVID-QU-Ex is a curated dataset for COVID-19 diagnosis from chest X-rays, containing labeled images annotated by clinical experts. The database contains a total 3487 images of which 423 are COVID-19, 1485 are viral pneumonia, and 1579 are labeled as normal chest X-ray images.

## 5.2 Experimental Setup

We structure our experimental pipeline into two primary stages: (i) visual and textual concept generation, and (ii) concept bottleneck model training and evaluation. This modular setup ensures that concept discovery and interpretability learning are decoupled and can be independently analyzed. The first stage of the pipeline is responsible for identifying interpretable visual regions and semantic descriptions associated with each input sample. This step is completely model-agnostic and dataset-specific.

**Visual Concept Generation:** For each input image  $x$ , we generate a set of  $M$  segmentation masks using SAM or MedSAM. Each mask  $s_i \in \{0, 1\}^{H \times W}$  isolates a distinct region of the image, representing a candidate visual concept. To efficiently extract embeddings, we form a batch containing:

- The full image  $x$
- The  $M$  masked variants  $x \odot s_1, \dots, x \odot s_M$

The resulting embeddings are arranged as:

$$E = [f_{\text{clip}}(x), f_{\text{clip}}(x \odot s_1), \dots, f_{\text{clip}}(x \odot s_M)] \in \mathbb{R}^{1+M \times d} \quad (5.1)$$

During experimentation phases, we parse this batch by separating the first embedding (representing the whole image) from the subsequent  $M$  segment-level embeddings as needed:

$$z = f_{\text{clip}}(x) \quad (5.2)$$

$$P^{\text{seg}} = \{f_{\text{clip}}(x \odot s_i)\}_{i=1}^M \quad (5.3)$$

**Textual Concept Generation.** For each diagnostic label in our dataset, we generate a list of descriptive textual phrases using LLMs with a structured prompt (see Section 4.2). These phrases are designed to reflect radiologically relevant, observable patterns. Each phrase is embedded using the frozen CLIP text encoder:

$$P^{\text{text}} = \left\{ f_{\text{clip}}^{\text{text}}(c_j) \right\}_{j=1}^C, \quad (5.4)$$

where  $C$  is the number of textual concepts associated with the label.

These visual and textual representations are then aligned during concept bottleneck training, as described in Section 4.3.2.

### Concept Bottleneck Model Training

In the second stage, we project full-image embeddings into the concept space and train a sparse final classifier. This stage uses precomputed features and concepts from the first stage. We use pre-extracted CLIP image embeddings and train a projection layer to map these embeddings into a concept space of dimensionality  $C$ . The projection is optimized using the combined loss function defined in Section 4.3.2, including:

- Output similarity loss ( $\mathcal{L}_{\text{out}}$ ),
- Local concept alignment loss ( $\mathcal{L}_{\text{cl}}$ ),
- Aggregated feature similarity loss ( $\mathcal{L}_{\text{agg}}$ ).

Training is performed using mini-batch SGD over a fixed number of projection steps, with all backbone encoders frozen. The output of the projection layer is number of total visual concepts multiplied with number of total textual concepts. Once the projection weights are learned, we freeze them and train a sparse linear classifier and the sparsity level is monitored by computing the ratio of non-zero weights.

## 5.3 Evaluation of Visual Concept Generation

We evaluate our concept-based explainability framework across five datasets as mentioned. The evaluation focuses on visual concept attribution performance using Area Under the Curve (AUC) as the primary metric, computed across five independent runs for each dataset. The results are compared against the different baseline when available. The results in Table 5.2 demonstrate that our method either outperforms or is comparable to EAC across all tested datasets. On ImageNet and COCO, our framework slightly improves over EAC. On CheXpert, a more domain-specific dataset, we observe a consistent gain in mean AUC and lower variance indicating stability.

Table 5.2: Mean AUC (%) and standard deviation (%) for visual concept attribution across 5 runs.

Dataset	Model	Mean AUC	Std. Dev.
ImageNet	EAC [118]	54.41	$\pm 13.97$
	<b>Our approach</b>	<b>58.00</b>	$\pm 11.05$
COCO	EAC [118]	46.68	$\pm 16.15$
	<b>Our approach</b>	<b>47.36</b>	$\pm 16.96$
CheXpert	EAC [118]	39.94	$\pm 7.28$
	<b>Our approach</b>	<b>41.04</b>	$\pm 8.94$

### 5.3.1 Impact of Domain-Specific Models

To explore the effect of domain adaptation, we incorporated MedSAM—a medical instance segmentation model—into our pipeline. Table 5.3 compares the attribution performance with and without MedSAM on the CheXpert dataset. Using MedSAM, AUC score improves to 86.62%, highlighting that segmentation precision correlates with concept attribution effectiveness. Stability of the model across five runs is also highly increased indicating that domain specific adaptation is needed at this stage to accurately curate visual concepts.

Table 5.3: Performance impact of MedSAM on CheXpert (5 runs).

Model	Mean AUC	Std. Dev.
EAC [118]	39.94	$\pm 7.28$
Our approach (SAM)	<b>41.04</b>	$\pm 8.94$
MedSAM + Our approach	<b>86.62</b>	$\pm 1.02$

### 5.3.2 Performance Across Medical Datasets

In Table 5.4, we report the final attribution AUC across all medical datasets supported by our framework. Our approach achieves strong performance on both MIMIC and COVID-QU, with AUCs exceeding 81.59%, demonstrating generalizability across different radiological image and disease distributions.

Table 5.4: Faithfulness - AUC scores with standard deviation across medical datasets (5 runs).

Dataset	Insertion		Deletion	
	AUC	Std. Dev.	AUC	Std. Dev.
CheXpert	<b>0.8662</b>	0.0102	<b>0.8497</b>	0.0206
MIMIC-CXR	0.8377	0.0082	0.8369	0.0169
COVID-QU	0.8159	0.0140	0.8232	0.0115

These findings confirm that our framework not only scales to multiple datasets but also achieves robust concept-level attribution without requiring extensive model re-training or handcrafted concept filtering.



### 5.3.3 Qualitative Interpretability

In addition to quantitative results, we qualitatively assess the interpretability of our visual concept explanations using SHAP-based overlays. Figure 5.1 and Figure 5.2 show sample outputs from our concept attribution pipeline on chest X-ray images, each overlaid with identified visual concepts (segmentation masks) and their respective contributions (SHAP scores). In other words, the segments represent automatically discovered concepts whose importance values are derived via Shapley value approximations.

These overlays offer a human-understandable explanation by highlighting the anatomical regions (e.g., cardiac silhouette, mediastinum, lung fields) that most influenced the model’s decision. Higher SHAP scores indicate stronger contributions toward the final class prediction, providing a granular and spatially localized rationale. For instance, in the case of *Cardiomegaly* (Figure 5.1a), the model attributes higher importance to the cardiac area (0.17), consistent with clinical expectations.

This also demonstrates that our method can distinguish between subtle radiological findings such as *Cardiomegaly* and *Enlarged Cardiomediastinum*, as seen in Figure 5.1(top left and bottom left respectively). This ability to differentiate overlapping thoracic pathologies validates the effectiveness of our concept-based decomposition in understanding clinical semantics.

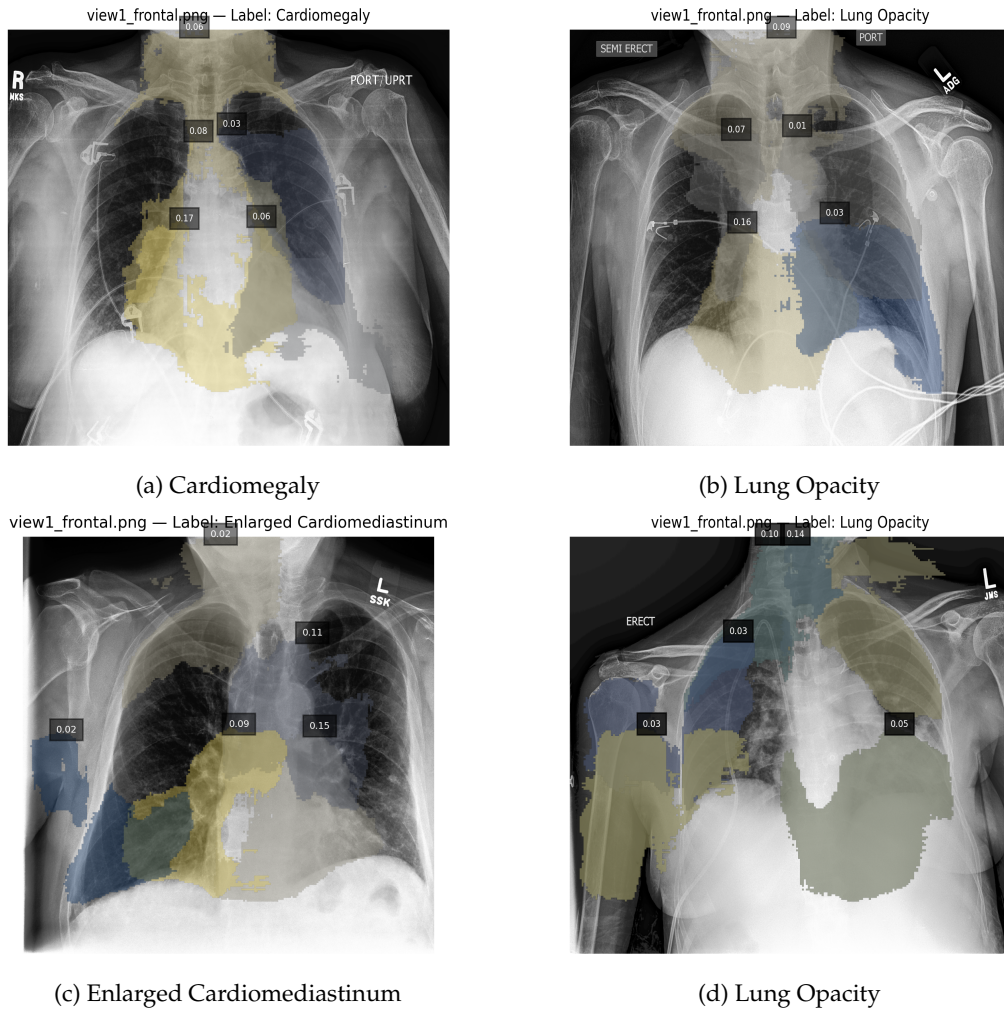


Figure 5.1: Extracted visual concept attributions on frontal chest X-rays. Their SHAP values overlaid on segmented regions reflect their importance in predicting respective findings.

However, we also note that the model struggled to extract concepts from Chest X-rays with lateral view (shown in figure 5.2) and often assigns the dark regions on the side with higher importance. The segmentation model also struggles to extract semantically aligned segments from these images, causing the VCG to incorrectly learn importances in such scenerios.

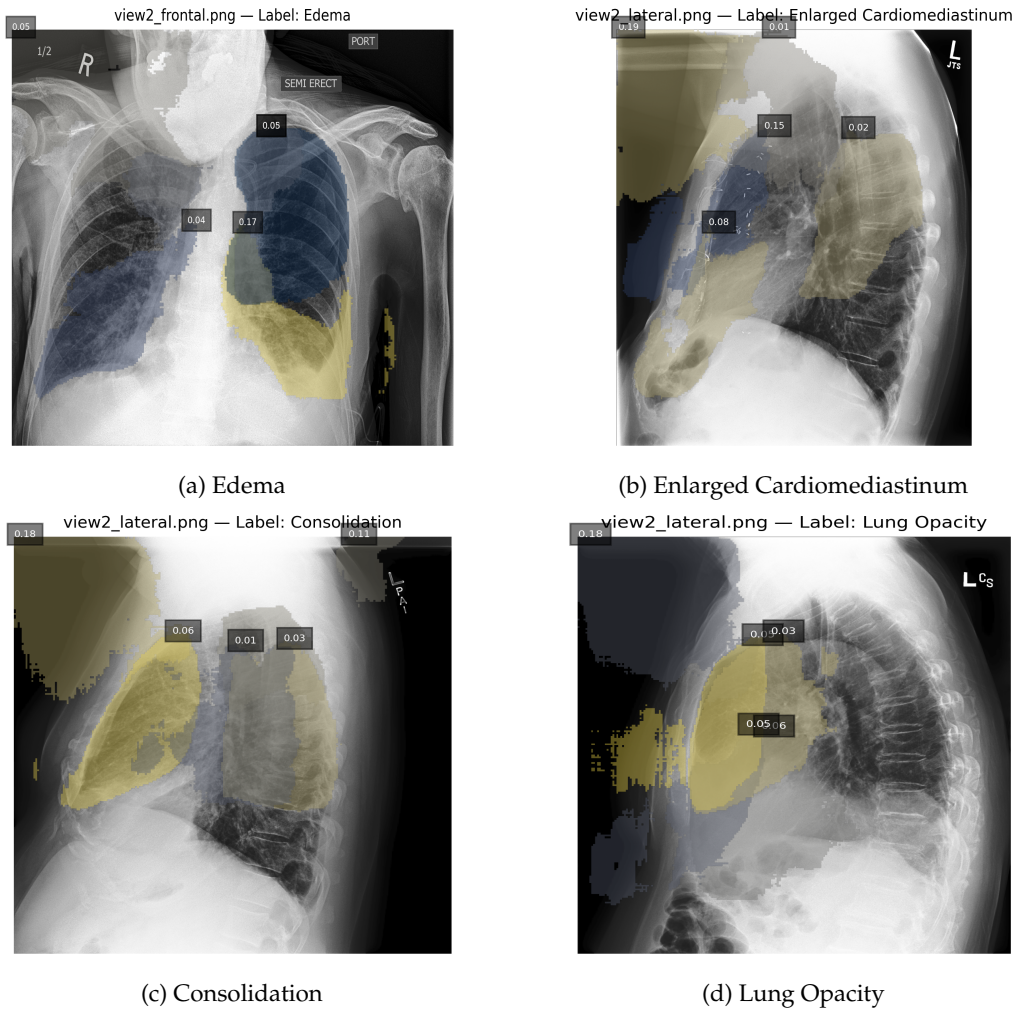


Figure 5.2: Visual Concept based interpretability across different viewpoints and findings

The qualitative understanding of our approach confirms that our visualizations align with the principles of concept-based explainability; as it offers explanations that are faithful, interpretable, and localized. They reinforce the diagnostic potential of the method, especially in domains like radiology where spatial reasoning and exact localization is critical. Overall, this qualitative evidence complements our AUC-based metrics and supports the adoption of interpretable, concept-guided decision pipelines in clinical AI systems.

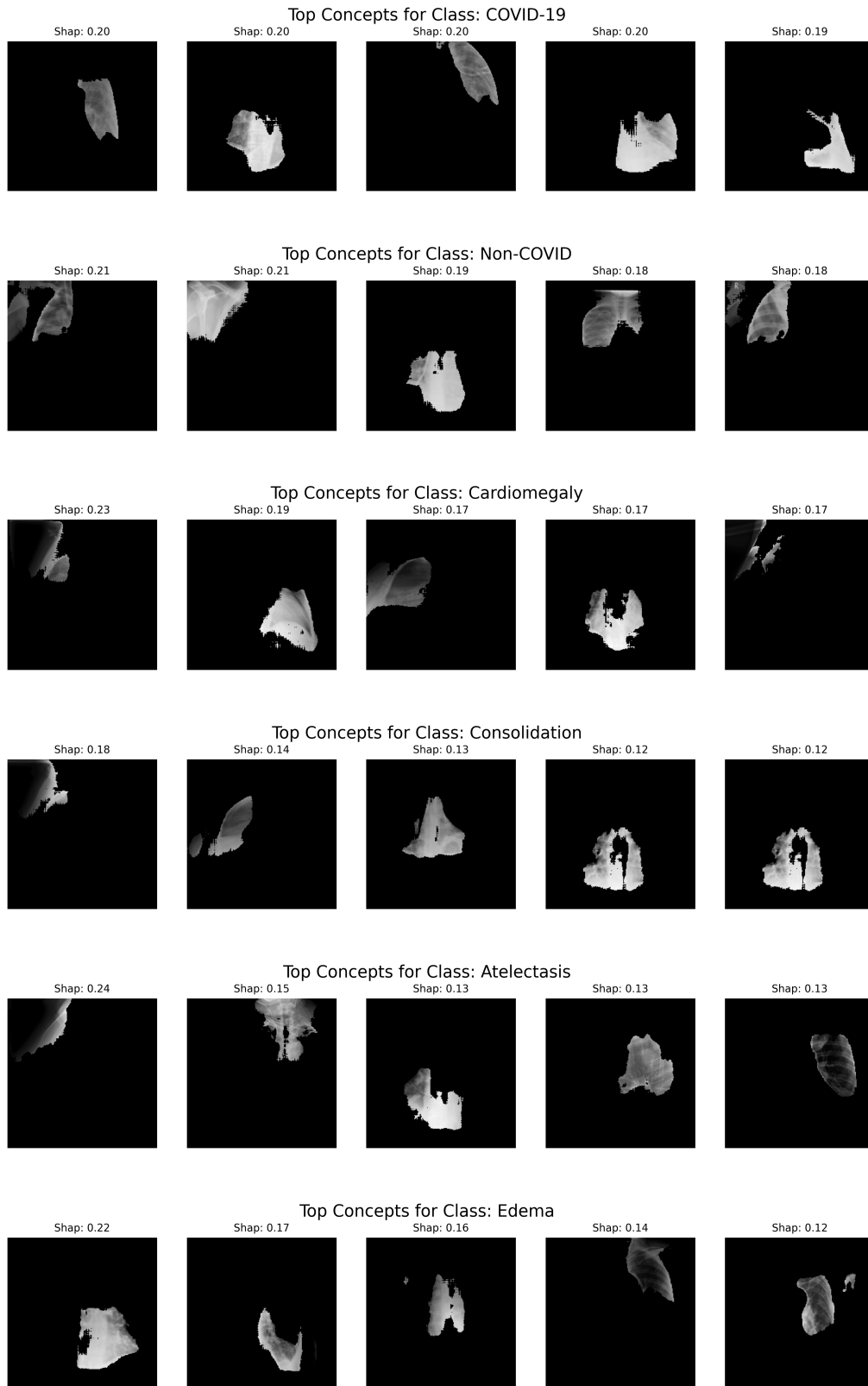


Figure 5.3: Top 5 concepts for each class across different Chest X-ray Samples from COVID-QU [29], MIMIC-CXR [56] and CheXpert [21]. In this approach, concept discovery is highly dependant on the segmentation model used, thus concepts may lack completeness in clinical semantics. The average accuracy across all datasets is 83.42%; indicating the potential of visually interpretable classification using this approach and therefore concept discovery can be highly improved by using a more accurate medical segmentation model.

## 5.4 Evaluation of Concept Bottleneck Model (CBM)

To contextualize our method, we compare against prior Concept Bottleneck Model (CBM) approaches in medical imaging. Table 5.5 summarizes accuracy, supervision requirements, and concept source across the MIMIC-CXR, CheXpert, and COVID-QU datasets. Our method demonstrates competitive performance on COVID-QU and MIMIC while using no manual labels and maintaining sparsity.

Table 5.5: Comparison of concept bottleneck models in medical imaging. Our method uses no manual labels and achieves competitive accuracy.

Model	Dataset	Accuracy	Manual	Sparse
Label Free CBM [91]	MIMIC	62.40%	✗	✓
Interpretable-CXR [132]	MIMIC	63.27%	✓	✗
Ours	MIMIC	68.35%	✗	✓
Label Free CBM [91]	COVID-QU	72.23%	✗	✓
Interpretable-CXR [132]	COVID-QU	78.00%	✓	✗
CBM-RAG [5]	COVID-QU	81.00%	✗	✗
Ours	COVID-QU	<b>83.43%</b>	✗	✓
Label Free CBM [91]	CheXpert	43.60%	✗	✓
Ours	CheXpert	43.35%	✗	✓

While our method underperforms slightly on CheXpert compared to prior CBMs, it achieves a new state-of-the-art on COVID-QU and shows substantial gains on MIMIC; without any manual supervision. These results highlight the potential of leveraging large-scale pretrained vision-language models and automated concept extraction pipelines for interpretable and scalable medical decision systems.

### 5.4.1 Evaluation on MIMIC-CXR

To further analyze the effectiveness of different modalities and design choices within our framework, we perform an ablation study on the MIMIC dataset. Results are reported in terms of validation accuracy and sparsity. Sparsity here serves as a score for model interpretability: lower values indicate fewer concepts used for prediction.

Table 5.6: Ablation on modality combinations, similarity cutoff, and sparsity vs. performance tradeoffs for MIMIC. All models use CLIP ViT-B/32 as backbone.

Textual	Visual	Sim Cutoff	Train Acc	Val Acc	Val Sim	Sparsity (%)
Y	Y	0.2	0.6446	0.6835	0.6525	85.71
Y	Y	0.35	0.7008	0.6392	0.6601	64.88
Y	N	0.35	<b>0.8586</b>	<b>0.7910</b>	<b>0.6717</b>	<b>60.19</b>

This analysis reveals several key insights. First, the higher similarity cutoffs encourage sparsity when only textual concepts are learnt in the projection layer. However, this is too restrictive when both textual and visual concepts are to be aligned. Lowering the cutoff, in other words, allowing slightly less similar textual and visual concepts allows the model retain concepts that are important for the classification task itself. As seen in

table 5.6, the model achieved 68.35% accuracy when the similarity cutoff was lowered to accommodate both textual and visual concepts.

### 5.4.2 Evaluation on CheXPert

We further investigate the performance of our Concept Bottleneck Model (CBM) across multiple model architectures and segmentation configurations for CheXpert dataset as it underperforms. Table 5.7 summarizes the CBM performance on the CheXpert dataset using different backbones and loss weights.

Table 5.7: CBM results on CheXpert . All models use max-pooling for visual concept aggregation.

Target Model	CLIP	$\lambda_{\text{cl}}$	Batch Size	Val Acc	Sparsity
CheXAgent	ViT-B/16	5	128	0.5800	0.1210
CheXAgent	ViT-B/16	5	64	<b>0.6100</b>	0.1297
ViT-XRay	ViT-B/16	2	128	0.4900	0.6294
ViT-XRay	ViT-B/16	2	64	0.5899	0.1427
CLIP-RN50	CLIP-RN50	1	64	0.6974	0.1431
CLIP-RN50	CLIP-RN50	2	128	<b>0.7560</b>	0.4418

We observe that when target model and embedding models are same, the explanations are more accurate, for example, CLIP-RN50 reaching 75.6% validation accuracy while maintaining reasonable sparsity (0.44). CheXAgent[27] models achieve moderate accuracy (57–61%) with notably moderate sparsity (0.12–0.13), suggesting more compact decision logic. Models using ViT-XRay (CLIP trained on chest X-rays) exhibit the highest sparsity but lower accuracy, indicating underfitting or overcompression.

We also note that, the removal of visual concepts while retaining only textual supervision leads to the highest validation accuracy and similarity alignment. This suggests that textual concept supervision alone can provide sufficient semantic abstraction when grounded in strong pretrained vision-language encoders like CLIP. Even though visual concepts help structure the concept space, they may introduce noise or redundancy due to dark or abstract regions in present in chest X-rays.

### 5.4.3 Qualitative Interpretability

In this subsection, we analyze how sparsity and concept attribution contribute to interpretability, and present representative case studies to illustrate explanation behaviour. Our final linear classifier exhibits significant sparsity (upto 85%) across multiple backbones. This allows us to isolate which concepts directly influence a model’s output by inspecting the non-zero weights in  $W_g$ . Each non-zero entry in  $W_g$  corresponds to a specific visual-textual concept, which can be interpreted by:

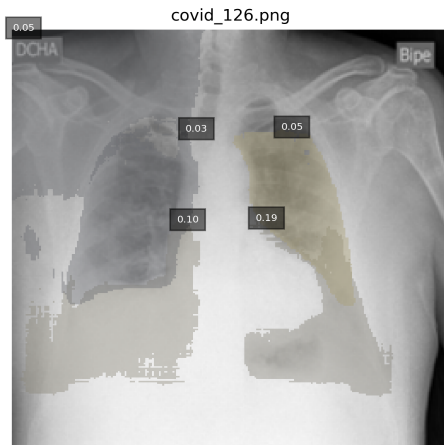
- Identifying the textual concept  $c_j$  it corresponds to
- Mapping it to the most similar visual concept  $s_i$  using cosine similarity

This mapping enables clinicians to inspect both spatial features and clinically relevant concepts that contributed most to the model’s decision.

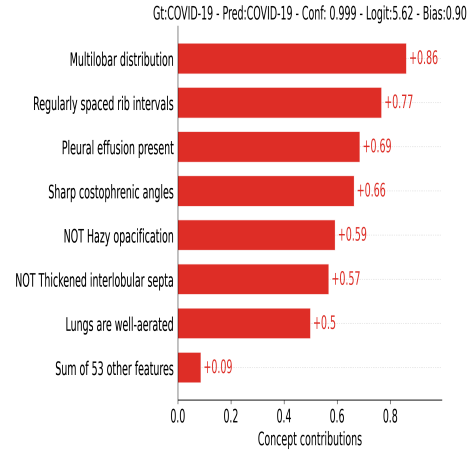
**Example: COVID Classification** In a representative prediction from the COVID-QU dataset, our model predicted **COVID-19** with a confidence of 0.967. The following concepts were most influential in the decision, as shown in Figure 5.4(d) textual concept contributions:

- **Increased interstitial markings** — strong positive contribution (+1.71); localized in interstitial regions indicating pathology.
- **Multilobar distribution** — also positively weighted (+1.41); signals diffuse abnormality across multiple lobes.
- **Pleural effusion present** — supports diagnosis (+1.13); typically associated with severe respiratory conditions.
- **Peripheral ground-glass opacities** — negatively weighted (−1.17); potentially contradictory or weakly matched in visual region.
- **Hazy opacification, Unremarkable soft tissues and bones, and Air bronchograms within consolidation** — moderate negative contributions; may indicate competing or less relevant patterns.

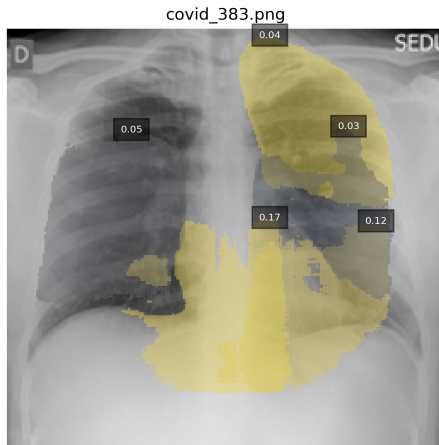
These textual concept-level attributions provide global insight into the model’s reasoning while visual concepts provide localized insights, helping to validate correct predictions, quickly identify biased predictions and the influence of clinically relevant markers.



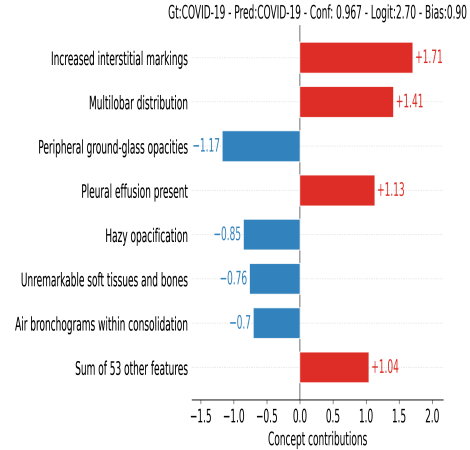
(a) Visual Concept Contributions



(b) Textual Concept Contributions



(c) Visual Concept Contributions



(d) Textual Concept Contributions

Figure 5.4: Visual and Textual concept attributions on Sample chest X-rays. The SHAP values overlaid on segmented regions reflect their importance in predicting respective findings. The Concept scores indicate the weight of the learned sparse layer in CBM architecture.

Each of these concepts corresponds to an active mask in the input, allowing clinicians to verify model rationale visually. While interpretability is improved, we also observe failure cases:

- Over-segmentation leads to redundant or irrelevant concepts
- Visual masks sometimes align with dark/gray regions or anatomical structures not relevant to the diagnostic task
- Textual concepts may be overly general (not distinctive enough) and appear spuriously

Such artifacts motivate future improvements in domain-specific filtering and concept consolidation.



#### 5.4.4 Global Concept Attribution

To gain insight into the global decision-making patterns learned by the Concept Bottleneck Model (CBM), we visualize the weights of the final linear classifier layer as Sankey diagrams. Each diagram illustrates the relative influence (positive or negative) of selected visual concepts on a target class, derived directly from the sparsely trained CBM weights. This complements the local visual concept attributions presented in section 5.3.3.

Figure 5.5 shows the concept contributions for *Cardiomegaly*, where influential concepts include “Ill-defined lung opacity” and “Mass with spiculated margins,” while the absence of lobulated or calcified lesions negatively contributes to the prediction. Such interpretability provides assistive clinical rationale behind the model’s outputs.

Concept Contributions to: Cardiomegaly

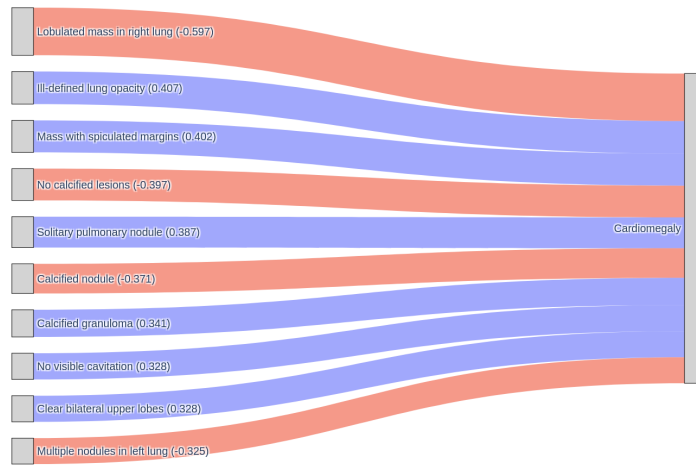


Figure 5.5: Learned concept contributions for **Cardiomegaly**. Positive contributions are shown in blue, negative in red. Weights are taken from the sparse linear classifier in the CBM.

We further visualize concept importance for additional classes across the MIMIC-CXR and COVID-QU datasets. These include *Lung Opacity* (Figure 5.6), *COVID-19* (Figure 5.7), and *Non-COVID vs. Normal* (Figure 5.8). The diagrams reveal interpretable structure in the learned weights—e.g., in the COVID-19 class, concepts like “Solitary circumscribed nodule” and “Left lung cavitation” drive the prediction, while absence of features (e.g., “No visible cavitation”) serve as counter-evidence. These global summaries not only validate that the learned representations align with radiological reasoning but also highlight the utility of concept sparsity for transparent model inspection.

Concept Contributions to: Lung Opacity

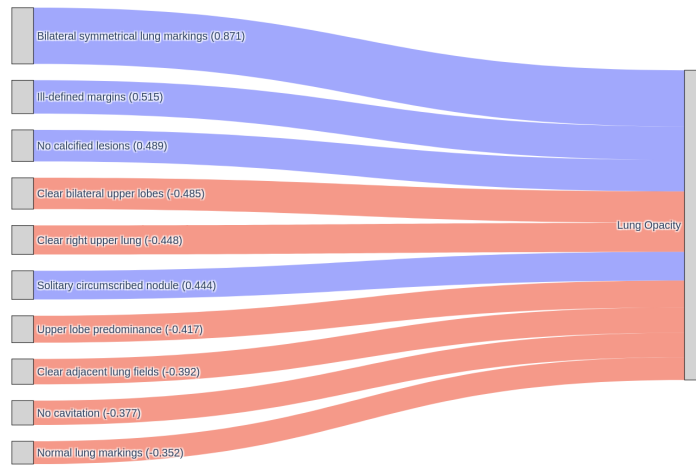


Figure 5.6: Concept contributions to **Lung Opacity**. CBM captures fine-grained inter-concept effects including suppression by clear findings.

Concept Contributions to: COVID-19

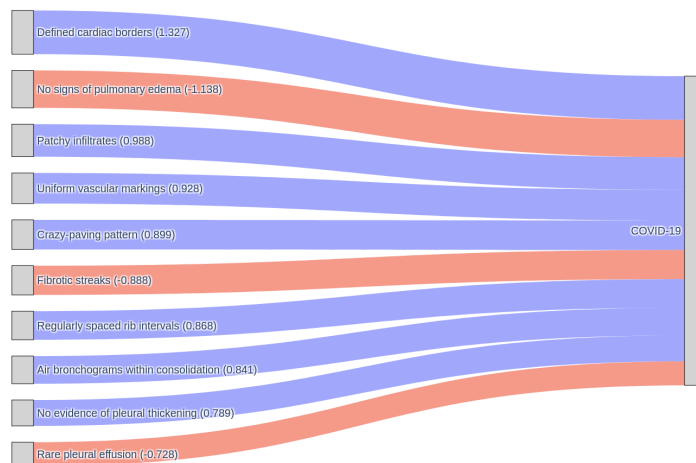
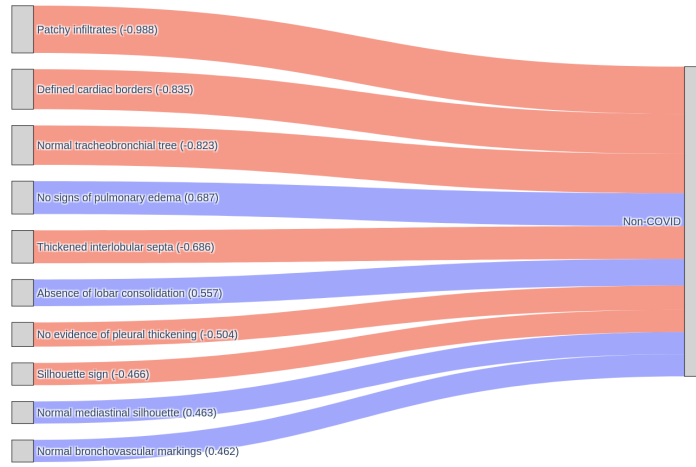


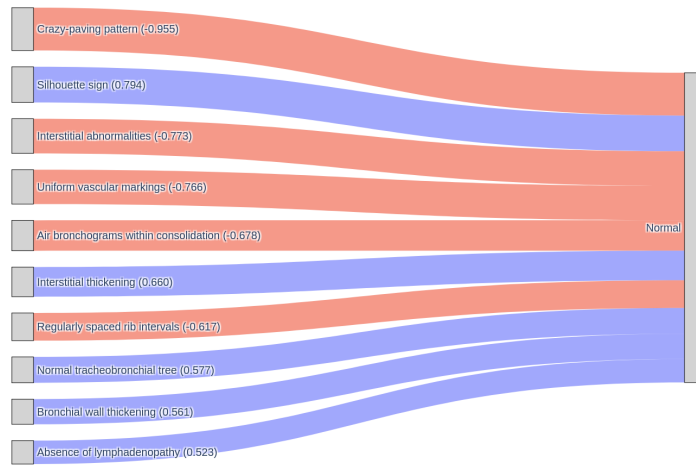
Figure 5.7: Learned concept contributions to **COVID-19**. Findings like “Left lung cavitation” and “Normal mediastinal contour” contribute positively.

Concept Contributions to: Non-COVID



(a) Non-COVID

Concept Contributions to: Normal



(b) Normal

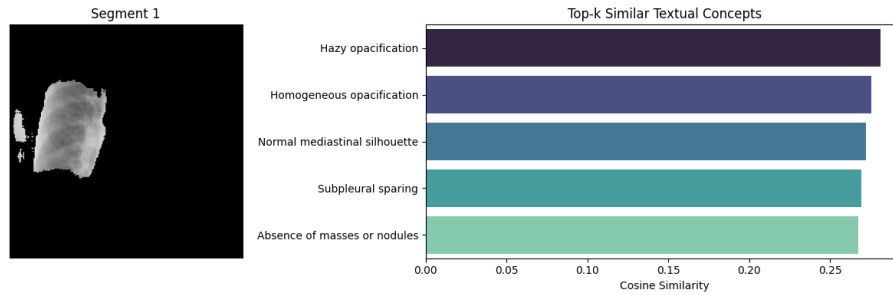
Figure 5.8: Concept-level decision patterns for **Non-COVID** and **Normal** classes. The model learns to suppress abnormal features in the Normal class.

These global visualizations underscore the interpretability advantage of sparse concept classifiers. They offer a holistic view of how individual concepts drive or counter specific

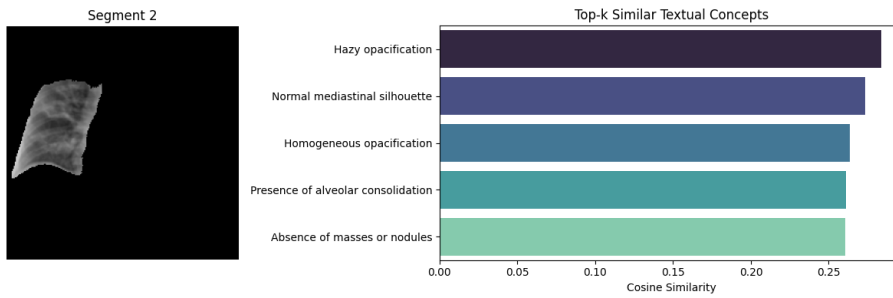
predictions, providing transparency beyond local feature attribution.

### 5.4.5 Visual-Textual Concept Alignment

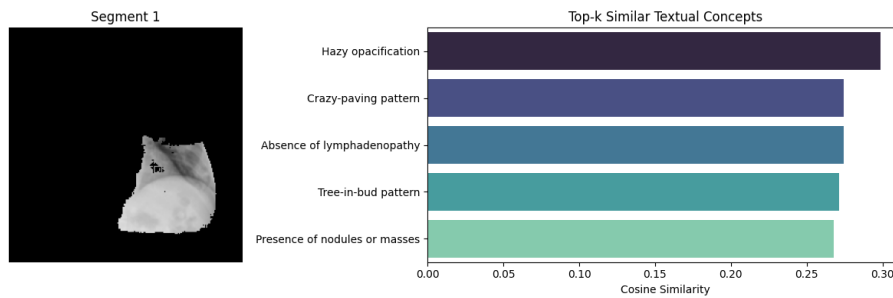
To demonstrate the semantic grounding of our visual concepts, we compute the cosine similarity between top-k visual concepts and the set of concepts filtered based on this alignment during CBM training. Figure 5.9 displays representative examples where each visual concept (segment) is paired with its top-5 most similar textual descriptions.



(a) Normal case: segment aligned with normality and absence of abnormalities.



(b) COVID-19 case: segment associated with consolidation and opacification.



(c) COVID-19 case: segment aligned with “Crazy-paving” and “Tree-in-bud” patterns.

Figure 5.9: Top-5 textual concepts semantically aligned with visual segments using cosine similarity. These associations confirm the medical interpretability of extracted visual concepts.

Across multiple cases, we observe high semantic consistency between the segmented

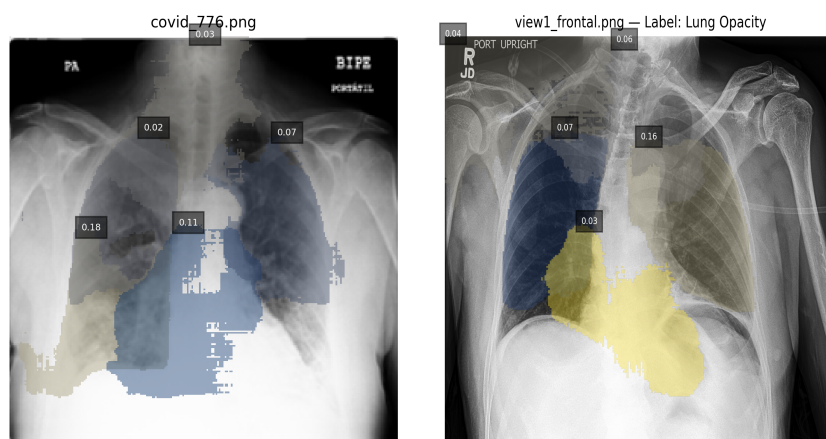
regions and their corresponding textual concepts. For instance, segments in COVID-19 cases align with descriptions such as “Hazy opacification,” “Crazy-paving pattern,” and “Presence of alveolar consolidation,” which are findings relevant to COVID-related pneumonia. Similarly, for a Normal case (Figure 5.9a), the aligned concepts include “Normal mediastinal silhouette” and “Absence of masses or nodules,” reinforcing the effectiveness of visual-textual alignment as a complementary interpretability strategy. These results demonstrate that the discovered visual concepts are not arbitrary or overfitted, but instead are aligned with clinically meaningful descriptors, thereby enhancing the transparency and trustworthiness of the concept bottleneck model pipeline.

## 5.5 Interpretability and Utility of Generated Explanations

Figure 5.10 demonstrates the interpretability benefits of our visual concept-based explanation framework applied to chest radiographs. Each of these showcases how our approach highlights that the model’s decision making is based on anatomically meaningful segments (using Shapley-based attribution values). This aligns with the aim of providing clinically understandable, and efficient explainability as emphasized in our approach.

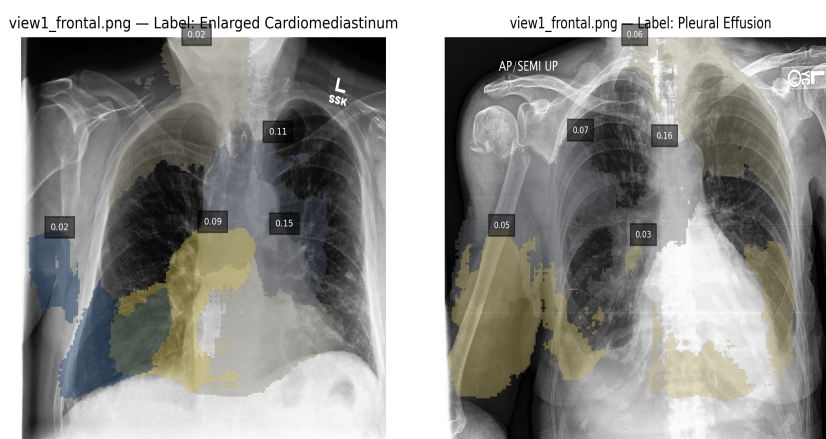
- **Figure 5.10(a)** highlights lower and perihilar lung zones contributing to a COVID-19 diagnosis. The model attributes high importance to these clinically relevant regions, reflecting proper feature learning of the vision-language model.
- **Figure 5.10(b)** shows accurate identification of mid-lung opacities for the Lung Opacity class. The two distinct highlighted zones reflect the model’s ability to localize pathological features effectively.
- **Figure 5.10(c)** focuses on the *Enlarged Cardiomediastinum* class. The model assigns high attribution scores to the central thoracic area, consistent with an enlarged cardiac silhouette, suggesting alignment with clinical reasoning.
- **Figure 5.10(d)** corresponds to *Pleural Effusion*. The model emphasizes basal lung regions and costophrenic angles, where effusion typically appears, thereby confirming diagnostic relevance.

These visualizations collectively validate the utility of concept-based explanations as they provide not only to understand model reasoning but also help establish trust in the model’s decision making is **anatomically grounded and semantically meaningful**. Overcoming the limitations of pixel or superpixel-level saliency maps, which do not clearly identify such segments or assign influence scores in predictions, our approach enhances interpretability by providing these information and building trust. On the other hand, these also enable expert clinicians to localize errors and identify model failure in cases of misclassification or inattentive focus; thereby increasing transparency.



(a) The model highlights important regions that contribute to COVID diagnosis.

(b) The model correctly assigns high importance to two zones to distinguish and diagnosis lung opacity.



(c) Visual segments highlight the high importance assigned to the enlarged cardiac area.

(d) Visual segments highlight the chest cavity and regions near the diaphragm important to identify pleural effusion.

Figure 5.10: Visually segmented concepts contribute towards higher interpretability and quick understanding of model decisions based on important regions. In lateral views, when model is confused or misidentifies regions, visual importances make it easily detectable.

Figure 5.11 presents concept-level attribution bar plots for two classification instances—one positive for COVID-19 and one negative (Non-COVID). For the COVID-19 positive case, the model heavily weights features such as *Pleural effusion present*, *Multilobar distribution*, and *Patchy infiltrates*. These are clinically established indicators of COVID-related pulmonary conditions. On the contrary, bilateral involvement or patchy infiltrates are concepts related to COVID-19 diagnosis and are assigned as a negative contribution despite the prediction being accurate; this may indicate that due to presence of such conditions in NON-COVID cases, the model doesn't rely on these to distinguish COVID cases. This alignment with radiological priors increases the interpretability of the model,

as the decision is grounded in familiar clinical patterns. In the Non-COVID case, the model attributes strong positive weight to opposite indicators of COVID-specific features, such as *NOT Pleural effusion present*, *NOT Reticular opacities*, and *NOT Uniform vascular markings*. These negate pathological findings typically associated with COVID-19, thus reinforcing the model’s classification decision in other pathological conditions.

By using clinically meaningful concepts (e.g., “Pleural effusion” or “Uniform vascular markings”), explanations are directly interpretable by radiologists, as opposed to abstract vector spaces or pixel saliency. The signed contributions easily clarifies whether each concept supports or opposes a particular diagnosis, offering easily understandable insight into model behavior. The relative magnitude of concept contributions highlights which factors were most influential in the final decision, not only bringing trust but possibly enabling targeted validation, correction or differential diagnosis. In case of incorrect predictions, concept contributions can be audited to identify over-reliance on non-discriminative or clinically less meaningful features, enabling diagnosis refinement or expert reevaluation. Thus, the concept-level framework offers a powerful tool for bridging machine predictions and expert decision-making, promoting actionable and trustworthy AI in clinical settings.

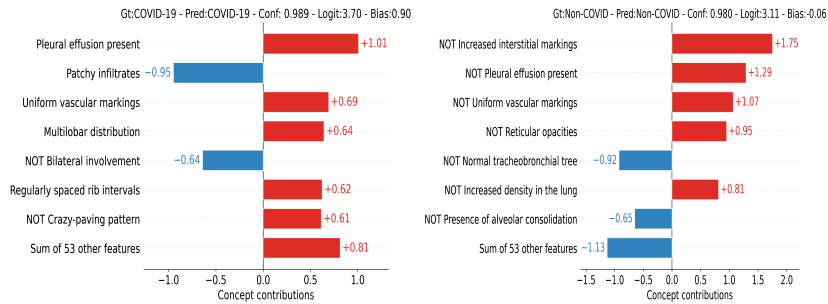


Figure 5.11: Concept contributions for COVID and Non-COVID disease diagnosis is distinct, aligning with clinically important factors such as pleural effusion being an indicator of the presence of COVID infection.

## 5.6 Case Studies and Limitations

Below, we examine two cases: one with ground truth *Lung Opacity* and another with *Cardiomegaly*. Each example includes both a textual concept barplot and an overlay of important regions on the chest X-ray.

### 5.6.1 Case 1: Lung Opacity

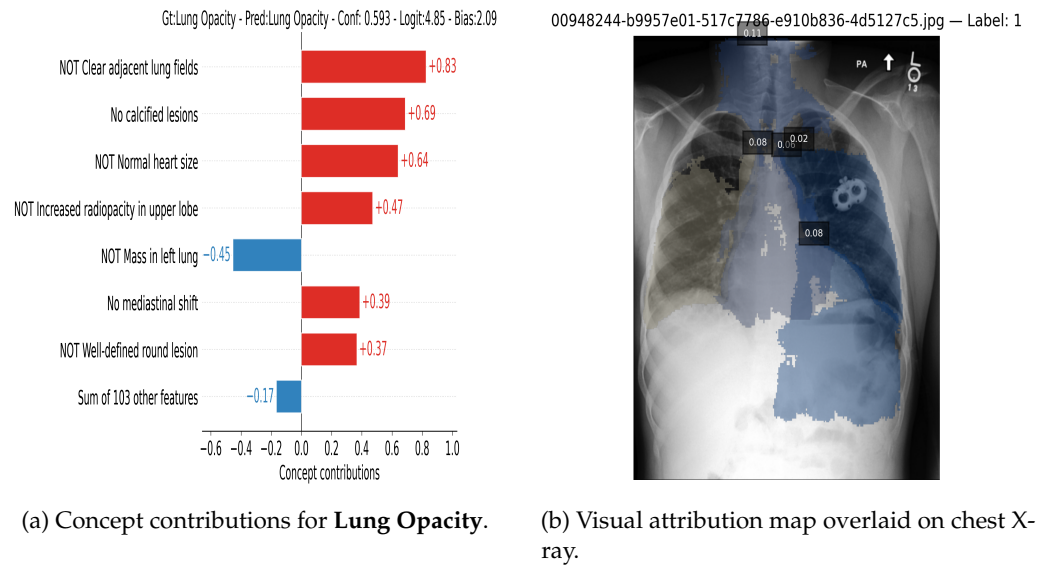
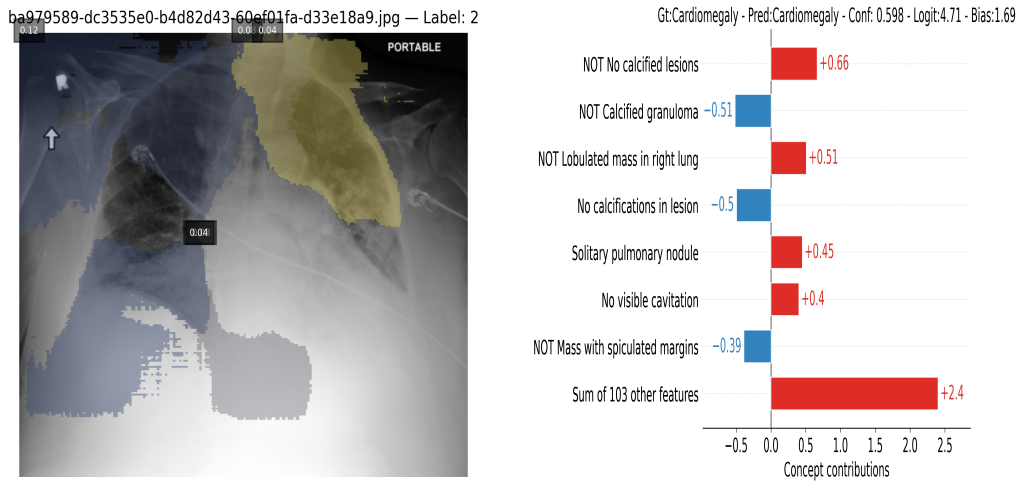


Figure 5.12: Explanation of the model's prediction for the **Lung Opacity** class (Confidence: 0.593).

The prediction for lung opacity (Figure 5.12a) is supported by strong positive contributions from concepts like *NOT Clear adjacent lung fields* (+0.83) and *No calcified lesions* (+0.69). The corresponding overlay in Figure 5.12b highlights multiple small, spatially dispersed regions across the lungs. However, the overall confidence is moderate (0.593), and the peak contribution values from individual segments are low (e.g., 0.08, 0.02), suggesting that no dominant visual region drove the prediction.



### 5.6.2 Case 2: Cardiomegaly



(a) Concept contributions for **Cardiomegaly**. (b) Visual attribution map overlaid on chest X-ray.

Figure 5.13: Explanation of the model's prediction for the **Cardiomegaly** class (Confidence: 0.598).

The cardiomegaly case shows that most of the predictive weight is carried by the aggregate of many small features (+2.4), while individual visual concepts contribute only moderately (e.g., *NOT No calcified lesions*: +0.66). Figure 5.13b shows dispersed regions with very low attribution scores (e.g., 0.04 or 0.03), primarily focused on the cardiac and upper mediastinal area.

### 5.6.3 Limitations

Although concept attribution and spatial overlays offer transparency, several limitations persist:

- **Low Contribution Scores:** The individual visual concept scores are relatively low in magnitude (mostly  $<0.1$ ), making it difficult to confidently attribute a specific region or feature to the prediction.
- **Diffuse Attribution Maps:** The highlighted regions are widespread and sometimes non-specific, reducing interpretability in clinical settings where localized pathology is critical (e.g., nodules, effusions). This is highly dependant on the segmentation model itself and not a limitation of the approach.
- **Over-reliance on Composite Features:** In the cardiomegaly case, a large proportion of the logit contribution (+2.4) comes from the sum of over 100 minor features. This suggests that the model lacks a dominant, easily explainable signal and instead relies on weak correlations spread across many features.
- **Multilabel Challenges:** These examples hint at the utility of the approach for multilabel tasks, where overlapping features might support multiple conditions

(e.g., lung opacity and cardiomegaly co-occurring). However, due to low concept fidelity and region specificity, the model may struggle to disentangle correlated classes or provide class-specific attributions.

- **Visual Concept Quality:** Some concepts (e.g., *No calcified lesions*, *No cavitation*) may not manifest visually in the X-ray or might require clinical correlation, limiting their standalone utility for image-based models.
- **Limited Confidence:** Both cases present predictions with confidence around 0.59, indicating model uncertainty. Explanations derived from such low-confidence outputs may be inherently unreliable.

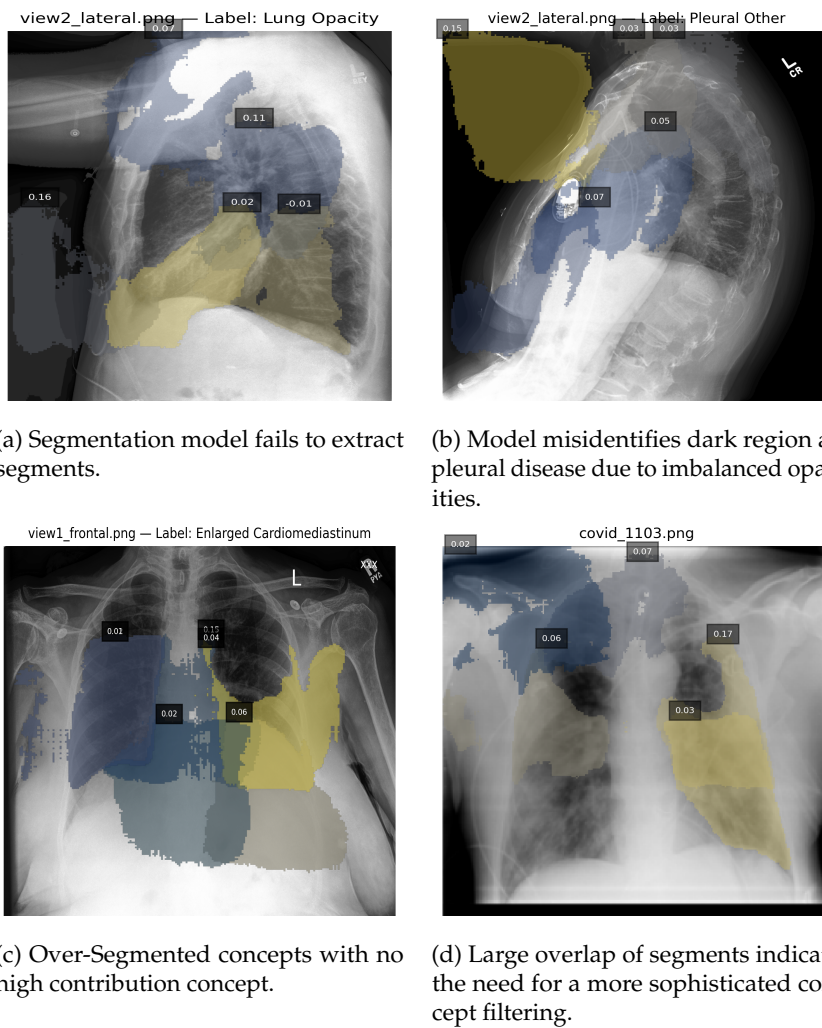


Figure 5.14: Limitations of Visual Segments and failure cases.

These examples illustrate the promise and limitations of this approach. While the model incorporates many semantically meaningful cues, the lack of dominant signals and the

diffuse nature of attribution maps suggest the need for more accurate segmentation models in visual concept extraction and curated textual concepts —especially for multilabel medical imaging tasks.

---

## Chapter 6

# Conclusion

We proposed a fully automated, modular framework for interpretable image classification based on concept bottlenecks. Integrating instance segmentation models with cross-modal embeddings and prompt-generated textual concepts, our system constructs a semantically grounded, sparse concept layer that facilitates transparent decision-making. Through surrogate modeling via the PIE architecture, we compute concept-level attributions efficiently while maintaining faithfulness. Empirical results across general and medical datasets show that our approach yields competitive classification accuracy while substantially improving interpretability. Our evaluation is mainly focused on the radiology domain spanning across publicly available datasets: CheXpert, MIMIC-CXR and COVID-QU. Notably, our method avoids manual annotations or expert-defined concepts, enhancing scalability. In MIMIC-CXR and COVID-QU, our framework maintains strong accuracy and transparency without relying on expert annotations or handcrafted ontologies. This allows the approach to be both scalable and generalizable, especially in medical settings where manual curation and expert intervention is costly or often infeasible. While concept bottlenecks perform well in structured settings with reliable labels, performance drops in label-sparse or noisy domains like CheXpert. Fine-grained segments may increase coverage but risk introducing redundant or noisy concepts and even confuse the model, while coarser abstractions may omit critical features. This underscores the need for adaptive segmentation and concept filtering mechanisms tuned to dataset properties. Despite the strengths of the pipeline, several limitations remain:

- **Concept Redundancy:** Retaining all segmentation masks without filtering leads to overlapping, low-utility, or semantically ambiguous concepts.
- **Lack of Semantic Labels:** Visual concepts are not explicitly always tied to textual or clinical terms, affecting understandability for end-users.
- **Segmentation Dependency:** Errors in the initial segmentation cascade through the concept bottleneck, impacting both interpretability and prediction.
- **Single-label Surrogates:** The surrogate model optimizes for a single predicted class, which constrains its effectiveness in multi-label or ambiguous scenarios which is often present in radiological diagnosis.

- **Concept Drift:** Concept activations sometimes align with vague or non-actionable features (e.g., “diffuse opacity”), limiting their diagnostic utility.
- **Contradictory Activations:** Without inter-concept constraints, conflicting evidence (e.g., “clear lungs” and “bilateral infiltrates”) may co-occur.
- **Contradictory Textual and Visual concepts:** Model may often confuse segmented concepts or dark region as cavities described in textual concepts, leading to misdiagnosis.

Involving humans-in-the-loop, particularly clinical experts at two key stages: textual concept curation and explanation validation can enhance trustworthiness and performance both. Structured feedback could also guide adaptive pruning and correction of mislabeled or spurious concepts [119].

## 6.1 Future Work

Building upon our current framework, future research can explore several avenues:

- **Semantic Concept Alignment:** Incorporating domain-specific knowledge bases to align learned concepts with established ontologies, enhancing interpretability.
- **Dynamic Concept Pruning:** Developing algorithms that dynamically prune irrelevant, conflicting or redundant concepts during inference to streamline explanations.
- **Cross-Domain Generalization:** Evaluating the transferability of learned concepts across different datasets to assess the robustness of the CBM framework.
- **Interactive Explanation Interfaces:** Designing user interfaces that allow end-users to interact with and query the model’s explanations, facilitating better understanding and trust.

These directions aim to enhance the scalability, adaptability, and user-friendliness of CBMs via allowing interactivity, promoting their adoption in real-world applications.

### 6.1.1 Integration with Large Language Models (LLMs)

Future research could explore the synergy between CBMs and LLMs, investigating how concept-based explanations can enhance the interpretability and controllability of large-scale language models. Current work can be extended to be predict token based on the curated concepts through an additional unsupervised learning layer proposed in Concept Bottleneck Large Language Models (CB-LLMs) [119]. Radiological reports can be analyzed to validate semantically relevant concepts via adopting cost-effective strategies [82] to further support the refinement of concepts in 4.2.

### 6.1.2 Spatially-Aware and Adaptive CBMs

Integrating spatial information into CBMs can enhance their interpretability, especially in vision tasks. The Spatially-Aware and Label-Free Concept Bottleneck Model (SALF-CBM) [14] introduces a framework that projects features into interpretable concept maps

without requiring human labels, producing high-quality spatial explanations. Adaptive CBM [30] on the other hand, re-examines the CBM framework specifically for medical image diagnostic tasks. Future work could explore combining spatial awareness with adaptive mechanisms to further improve the flexibility and interpretability of CBMs across diverse tasks.

### **6.1.3 Causal and Hierarchical Concept Modeling**

Understanding the causal relationships between concepts and predictions is crucial for trustworthy AI systems. Causally Reliable Concept Bottleneck Models (C2BMs) [34] aim to ensure that the concepts used by the model have a causal influence on the output, rather than merely being correlated. Additionally, incorporating hierarchical structures into concept modeling can alleviate information leakage issues by introducing label supervision in concept prediction and constructing hierarchical concept sets. Future research should delve into integrating causal inference techniques and hierarchical modeling within CBMs to enhance their robustness and interpretability.

### **6.1.4 Human-in-the-Loop and Editable CBMs**

Incorporating human feedback into CBMs can significantly improve their interpretability and trustworthiness. Editable Concept Bottleneck Models (ECBMs) [52] support various levels of data removal, allowing users to interactively edit concepts and observe the impact on model predictions. This interactive capability facilitates better understanding and debugging of model behavior. Future work should focus on developing user-friendly interfaces and methodologies for effective human-in-the-loop interactions with CBMs, enabling domain experts to guide and refine model explanations.

---

## Bibliography

- [1] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, and others. 2023. GPT-4 Technical Report. (2023). <https://api.semanticscholar.org/CorpusID:257532815>
- [2] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation. *Nature Machine Intelligence* 5, 11 (2023), 1006–1019. DOI:<http://dx.doi.org/10.1038/s42256-023-00711-8>
- [3] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers. *ArXiv abs/2402.05602* (2024). <https://api.semanticscholar.org/CorpusID:267547813>
- [4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:52938797>
- [5] Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. 2025. Towards Interpretable Radiology Report Generation via Concept Bottlenecks Using a Multi-agentic RAG. In *Advances in Information Retrieval*. Springer Nature Switzerland, 201–209. DOI:[http://dx.doi.org/10.1007/978-3-031-88714-7\\_18](http://dx.doi.org/10.1007/978-3-031-88714-7_18)
- [6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv abs/2204.14198* (2022). <https://api.semanticscholar.org/CorpusID:248476411>
- [7] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. XAI for Transformers: Better Explanations through Conservative Propagation. *ArXiv abs/2202.07304* (2022). <https://api.semanticscholar.org/CorpusID:246863594>
- [8] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018). <https://arxiv.org/abs/1806.08049>

- [9] Plamen P. Angelov, Eduardo Almeida Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11 (2021). <https://api.semanticscholar.org/CorpusID:236501382>
- [10] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 1–7. DOI:<http://dx.doi.org/10.18653/v1/W16-1601>
- [11] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10 (2015). <https://api.semanticscholar.org/CorpusID:9327892>
- [12] Shruthi Bannur, Stephanie L. Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel Coelho de Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria T. A. Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 15016–15027. <https://api.semanticscholar.org/CorpusID:255595629>
- [13] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus Müller, and Wojciech Samek. 2018. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *ArXiv abs/1807.03418* (2018). <https://api.semanticscholar.org/CorpusID:267800116>
- [14] Itay Benou and Tammy Riklin-Raviv. 2025. Show and Tell: Visually Explainable Deep Neural Nets via Spatially-Aware Concept Bottleneck Models. *ArXiv abs/2502.20134* (2025). <https://api.semanticscholar.org/CorpusID:276647844>
- [15] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truys, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. 2024. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. (2024). <https://arxiv.org/abs/2406.06512>
- [16] Angela Bonifati, Stefania Dumbrava, and Nicolas Mir. 2022. Hierarchical Clustering for Property Graph Schema Discovery. In *International Conference on Extending Database Technology*. <https://api.semanticscholar.org/CorpusID:247848516>
- [17] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901. DOI:<http://dx.doi.org/10.48550/arXiv.2005.14165>



- [18] Enrico Bunde, Daniel Eisenhardt, Daniel Sonntag, Hans-Jürgen Profitlich, and Christian Meske. 2023. *Giving DIANA More TIME – Guidance for the Design of XAI-Based Medical Decision Support Systems*. 107–122. DOI:[http://dx.doi.org/10.1007/978-3-031-32808-4\\_7](http://dx.doi.org/10.1007/978-3-031-32808-4_7)
- [19] Arianna Bunnell, Yannik Glaser, Dustin Valdez, Thomas Wolfgruber, Aleen Altamirano, Carol Zamora Gonz’alez, Brenda Y. Hernandez, Peter Sadowski, and John A. Shepherd. 2024. Learning a Clinically-Relevant Concept Bottleneck for Lesion Detection in Breast Ultrasound. *ArXiv abs/2407.00267* (2024). <https://api.semanticscholar.org/CorpusID:270869930>
- [20] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of Explainable AI Techniques in Healthcare. *Sensors* 23, 2 (2023), 634. DOI:<http://dx.doi.org/10.3390/s23020634>
- [21] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats. *arXiv preprint arXiv:2405.19538* (2024).
- [22] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. 2023. Interactive Concept Bottleneck Models. (2023). <https://arxiv.org/abs/2212.07430>
- [23] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. 347–356. DOI:<http://dx.doi.org/10.1109/ICCV48922.2021.00041>
- [24] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A Survey on Vision-Language Pre-training. *Machine Intelligence Research* 20, 1 (2023), 38–56. DOI:<http://dx.doi.org/10.1007/s11633-022-1369-5>
- [25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning UNiversal Image-TExt Representations. *ArXiv abs/1909.11740* (2019). <https://api.semanticscholar.org/CorpusID:202889174>
- [26] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision*. Springer, 104–120. <https://arxiv.org/abs/1909.11740>
- [27] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. 2024. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *ArXiv abs/2401.12208* (2024). <https://api.semanticscholar.org/CorpusID:267069358>
- [28] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:231802355>

- [29] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. 2020. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* 8 (2020), 132665–132676. DOI:<http://dx.doi.org/10.1109/ACCESS.2020.3010287>
- [30] Townim F. Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W. Verjans, and Zhibin Liao. 2024. AdaCBM: An Adaptive Concept Bottleneck Model for Explainable and Accurate Diagnosis. (2024). <https://arxiv.org/abs/2408.02001>
- [31] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. Multi-Head Attention: Collaborate Instead of Concatenate. *ArXiv abs/2006.16362* (2020). <https://api.semanticscholar.org/CorpusID:220265735>
- [32] Jonathan Crabbé and Mihaela van der Schaar. 2022. Concept Activation Regions: A Generalized Framework For Concept-Based Explanations. (2022). <https://arxiv.org/abs/2209.11222>
- [33] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv abs/2305.06500* (2023). <https://api.semanticscholar.org/CorpusID:258615266>
- [34] Giovanni de Felice, Arianna Casanova Flores, Francesco De Santis, Silvia Santini, Johannes Schneider, Pietro Barbiero, and Alberto Termine. 2025. Causally Reliable Concept Bottleneck Models. *ArXiv abs/2503.04363* (2025). <https://api.semanticscholar.org/CorpusID:276813286>
- [35] Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. 2014. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:218654665>
- [36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: a Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition* (06 2009), 248–255. DOI:<http://dx.doi.org/10.1109/CVPR.2009.5206848>
- [37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* (2019). DOI:<http://dx.doi.org/10.48550/arXiv.1810.04805>
- [38] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4443–4458. DOI:<http://dx.doi.org/10.18653/v1/2020.acl-main.408>
- [39] Jurgen Dieber and S. Kirrane. 2020. Why model why? Assessing the strengths and limitations of LIME. *ArXiv abs/2012.00093* (2020). <https://api.semanticscholar.org/CorpusID:227238867>

- [40] Chelluri Divakar, Ramanu Harsha, Kodali Radha, Dulipalla Venkata Rao, Nadakuditi Madhavi, and Thiruveedhula Bharadwaj. 2024. Explainable AI for CNN-LSTM Network in PCG-Based Valvular Heart Disease Diagnosis. *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (2024), 92–97. <https://api.semanticscholar.org/CorpusID:268613471>
- [41] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017). <https://arxiv.org/abs/1702.08608>
- [42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. DOI:<http://dx.doi.org/10.48550/arXiv.2010.11929>
- [43] Maximilian Dreyer, Reduan Achitbat, Wojciech Samek, and Sebastian Lapuschkin. 2023. Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations. *ArXiv abs/2311.16681* (2023). <https://api.semanticscholar.org/CorpusID:265466851>
- [44] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et al. 2024. The Llama 3 Herd of Models. *ArXiv abs/2407.21783* (2024). <https://api.semanticscholar.org/CorpusID:271571434>
- [45] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:184487319>
- [46] Shantanu Ghosh, K. Yu, and K. Batmanghelich. 2023. Distilling BlackBox to Interpretable models for Efficient Transfer Learning. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 14221 (2023), 628–638. <https://api.semanticscholar.org/CorpusID:258960495>
- [47] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (2018), 80–89. <https://api.semanticscholar.org/CorpusID:59600034>
- [48] Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence* 7 (2024). <https://api.semanticscholar.org/CorpusID:268249175>
- [49] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (1997), 1735–1780. <https://api.semanticscholar.org/CorpusID:1915014>
- [50] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C. Shen, George Shih, Ronald M. Summers, Yifan Peng, and Zhangyang Wang. 2022. *Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study*. Springer Nature Switzerland, 22–32. DOI:[http://dx.doi.org/10.1007/978-3-031-17027-0\\_3](http://dx.doi.org/10.1007/978-3-031-17027-0_3)

- [51] Lijie Hu, Tianhao Huang, Huanyi Xie, Chenyang Ren, Zhengyu Hu, Lu Yu, and Di Wang. 2024a. Semi-supervised Concept Bottleneck Models. *arXiv preprint arXiv:2406.18992* (2024).
- [52] Lijie Hu, Chenyang Ren, Zhengyu Hu, Cheng-Long Wang, and Di Wang. 2024b. Editable Concept Bottleneck Models. *ArXiv abs/2405.15476* (2024). <https://api.semanticscholar.org/CorpusID:270045383>
- [53] Tim Hulsén. 2023. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI* 4, 3 (2023), 578–588. DOI:<http://dx.doi.org/10.3390/ai4030034>
- [54] Johannes Jakubik, Sujit Roy, Christopher Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniel Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khalilagh, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Kiran Ganti, Kommy Weldemariam, and Rahul Ramachandran. 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. *ArXiv abs/2310.18660* (2023). <https://api.semanticscholar.org/CorpusID:264590307>
- [55] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zhifeng Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918* (2021). <https://arxiv.org/abs/2102.05918>
- [56] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *CoRR abs/1901.07042* (2019). <http://arxiv.org/abs/1901.07042>
- [57] Md Abdul Kadir, Gowtham Krishna Addluri, and Daniel Sonntag. 2023. Harmonizing Feature Attributions Across Deep Learning Architectures: Enhancing Interpretability and Consistency. *ArXiv abs/2307.02150* (2023). <https://api.semanticscholar.org/CorpusID:259341912>
- [58] Md Abdul Kadir, Hasan Md Tusfiqur Alam, Pascale Maul, Hans-Jürgen Profitlich, Moritz Wolf, and Daniel Sonntag. 2024. Modular Deep Active Learning Framework for Image Annotation: A Technical Report for the Ophthalmology-AI Project. (3 2024).
- [59] Md Abdul Kadir, Abdulrahman Mohamed Selim, Michael Barz, and Daniel Sonntag. 2023a. A User Interface for Explaining Machine Learning Model Explanations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 59–63. DOI:<http://dx.doi.org/10.1145/3581754.3584131>

- [60] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. 2023b. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)* (2023-01-01). IEEE, 000111–000124. [https://www.dfki.de/fileadmin/user\\_upload/import/14708\\_XAI\\_Evaluation\\_Metrics\\_\\_Taxonomies\\_\\_Concepts\\_and\\_Applications\\_\\_INES\\_2023\\_-7.pdf](https://www.dfki.de/fileadmin/user_upload/import/14708_XAI_Evaluation_Metrics__Taxonomies__Concepts_and_Applications__INES_2023_-7.pdf)<https://ieeexplore.ieee.org/abstract/document/10297629>
- [61] Md Abdul Kadir, Fabrizio Nunnari, and Daniel Sonntag. 2023c. Fine-tuning of explainable CNNs for skin lesion classification based on dermatologists’ feedback towards increasing trust. (2023). [https://www.dfki.de/fileadmin/user\\_upload/import/14709\\_2304.01399.pdf](https://www.dfki.de/fileadmin/user_upload/import/14709_2304.01399.pdf)<https://arxiv.org/abs/2304.01399>
- [62] Mert Keser, Gesina Schwalbe, Azarm Nowzad, and Alois Knoll. 2023. Interpretable Model-Agnostic Plausibility Verification for 2D Object Detectors Using Domain-Invariant Concept Bottleneck Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3891–3900. DOI:<http://dx.doi.org/10.1109/CVPRW59228.2023.00403>
- [63] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). (2018). <https://arxiv.org/abs/1711.11279>
- [64] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sung-Hoon Yoon. 2023a. Probabilistic Concept Bottleneck Models. *ArXiv* abs/2306.01574 (2023). <https://api.semanticscholar.org/CorpusID:259063823>
- [65] Hyunjoong Kim, Wei-Yin Loh, Yu-Shan Shih, and Probal Chaudhuri. 2007. Visualizable and interpretable regression models with good prediction power. *IIE Transactions* 39 (2007), 565 – 579. <https://api.semanticscholar.org/CorpusID:5843968>
- [66] In-Ho Kim, Jongha Kim, Joon-Young Choi, and Hyunwoo J. Kim. 2023b. Concept Bottleneck with Visual Concept Filtering for Explainable Medical Image Classification. *ArXiv* abs/2308.11920 (2023). <https://api.semanticscholar.org/CorpusID:261076347>
- [67] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 3992–4003. <https://api.semanticscholar.org/CorpusID:257952310>
- [68] Patrick Knab, Sascha Marton, and Christian Bartelt. 2024. DSEG-LIME: Improving Image Explanation by Hierarchical Data-Driven Segmentation. <https://api.semanticscholar.org/CorpusID:268363283>
- [69] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. (2020). <https://arxiv.org/abs/2007.04612>

- [70] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. DOI:<http://dx.doi.org/10.1109/5.726791>
- [71] Robert Leist, Hans-Jürgen Profitlich, Tim Hunsicker, and Daniel Sonntag. 2025b. Towards Trustable Clinical Decision Support Systems: A User Study with Ophthalmologists. In *IUI '25: Proceedings of the 30th International Conference on Intelligent User Interfaces. International Conference on Intelligent User Interfaces (IUI-2025), March 24-27, Cagliari, Italy*. Association for Computing Machinery.
- [72] Robert Leist, Hans-Jürgen Profitlich, Tim Hunsicker, and Daniel Sonntag. 2025c. Towards Trustable Intelligent Clinical Decision Support Systems: A User Study with Ophthalmologists. 1470–1484. DOI:<http://dx.doi.org/10.1145/3708359.3712136>
- [73] Robert Leist, Hans-Jürgen Profitlich, and Daniel Sonntag. 2025a. An AI-driven Clinical Decision Support System for the Treatment of Diabetic Retinopathy and Age-related Macular Degeneration. In *Joint Proceedings of the ACM IUI Workshops 2025. Workshop on Intelligent and Interactive Health User Interfaces (HealthIUI-2025), befindet sich IUI-2025, March 24-24, Cagliari, Italy*. Association of Computing Machinery, Association of Computing Machinery.
- [74] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *NeurIPS Datasets and Benchmarks Track*. <https://arxiv.org/abs/2306.00890>
- [75] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *International Conference on Machine Learning* (2023). DOI: <http://dx.doi.org/10.48550/arxiv.2301.12597>
- [76] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:246411402>
- [77] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:236034189>
- [78] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557* (2019). <https://arxiv.org/abs/1908.03557>
- [79] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2021. Grounded Language-Image Pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10955–10965. <https://api.semanticscholar.org/CorpusID:244920947>
- [80] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2022. Scaling Language-Image Pre-Training via Masking. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 23390–23400. <https://api.semanticscholar.org/CorpusID:254125280>

- [81] Siting Liang and Daniel Sonntag. 2025. Explainable Biomedical Claim Verification with Large Language Models. In *Joint Proceedings of the ACM IUI Workshops 2025. International Conference on Intelligent User Interfaces (IUI-2025), ACM IUI Workshops 2025, befindet sich IUI-2025, March 24-27, Cagliari, Italy*. Joint Proceedings of the ACM IUI Workshops 2025.
- [82] Siting Liang, Pablo Valdunciel Sánchez, and Daniel Sonntag. 2024. Optimizing Relation Extraction in Medical Texts through Active Learning: A Comparative Analysis of Trade-offs. In *Association for Computational Linguistics. Conference of the European Chapter of the Association for Computational Linguistics (EACL-2024), March 17-22, St. Julians, Malta*. ACL Anthology.
- [83] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. <https://api.semanticscholar.org/CorpusID:14113767>
- [84] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, Vol. 32. <https://arxiv.org/abs/1908.02265>
- [85] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. (2017). <https://api.semanticscholar.org/CorpusID:21889700>
- [86] Jun Ma and Bo Wang. 2023. Segment Anything in Medical Images. *ArXiv abs/2304.12306* (2023). <https://api.semanticscholar.org/CorpusID:258298289>
- [87] Eram Mahamud, Nafiz Fahad, Md Assaduzzaman, S.M. Zain, Kah Ong Michael Goh, and Md. Kishor Morol. 2024. An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning. *Decision Analytics Journal* (2024). <https://api.semanticscholar.org/CorpusID:270957593>
- [88] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. (2018). <https://arxiv.org/abs/1706.07269>
- [89] Duy MH Nguyen, Hoang Nguyen, Nghiem T Diep, Tan N Pham, Tri Cao, Binh T Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, and others. 2023. LVM-Med: Learning Large-Scale Self-Supervised Vision Models for Medical Imaging via Second-order Graph Matching. *arXiv preprint arXiv:2306.11925* (2023).
- [90] Fabrizio Nunnari and Daniel Sonntag. 2021. A Software Toolbox for Deploying Deep Learning Decision Support Systems with XAI Capabilities. In *Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '21)*. Association for Computing Machinery, New York, NY, USA, 44–49. DOI: <http://dx.doi.org/10.1145/3459926.3464753>
- [91] Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-Free Concept Bottleneck Models. *ArXiv abs/2304.06129* (2023). <https://api.semanticscholar.org/CorpusID:258107969>

- [92] Tuomas P. Oikarinen and Tsui-Wei Weng. 2022. CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. *ArXiv* abs/2204.10965 (2022). <https://api.semanticscholar.org/CorpusID:248376976>
- [93] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy T. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Huibin Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv.org* (2023). DOI:<http://dx.doi.org/10.48550/arxiv.2304.07193>
- [94] Konstantinos P. Panousis, Dino Ienco, and Diego Marcos. 2023. Sparse Linear Concept Discovery Models. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2023), 2759–2763. <https://api.semanticscholar.org/CorpusID:261049194>
- [95] Konstantinos P. Panousis, Dino Ienco, and Diego Marcos. 2024. Coarse-to-Fine Concept Bottleneck Models. (2024). <https://arxiv.org/abs/2310.02116>
- [96] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jirásková, Rickmer F. Braren, and Nassir Navab. 2023. Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis. *ArXiv* abs/2303.13391 (2023). <https://api.semanticscholar.org/CorpusID:257687236>
- [97] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. (2018). <https://arxiv.org/abs/1806.07421>
- [98] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. 8748–8763. <https://arxiv.org/abs/2103.00020>
- [99] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI Blog* (2018). [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) Technical Report.
- [100] Amy Rafferty, Rishi Ramaesh, and Ajitha Rajan. 2024. Transparent and Clinically Interpretable AI for Lung Cancer Detection in Chest X-Rays. *ArXiv* abs/2403.19444 (2024). <https://api.semanticscholar.org/CorpusID:268733344>
- [101] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *ArXiv* abs/2102.12092 (2021). <https://api.semanticscholar.org/CorpusID:232035663>
- [102] Navin Ranjan and Andreas E. Savakis. 2024. LRP-QViT: Mixed-Precision Vision Transformer Quantization via Layer-wise Relevance Propagation. *ArXiv* abs/2401.11243 (2024). <https://api.semanticscholar.org/CorpusID:267069149>



- [103] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. Model-Agnostic Interpretability of Machine Learning. (2016). <https://arxiv.org/abs/1606.05386>
- [104] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI: <http://dx.doi.org/10.1145/2939672.2939778>
- [105] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:3366554>
- [106] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673. DOI: <http://dx.doi.org/10.1109/TNNLS.2016.2599820>
- [107] Francesco De Santis, Philippe Bich, Gabriele Ciravegna, Pietro Barbiero, Danilo Giordano, and Tania Cerquitelli. 2024. Self-supervised Interpretable Concept-based Models for Text Classification. *ArXiv abs/2406.14335* (2024). <https://api.semanticscholar.org/CorpusID:270620407>
- [108] Ovi Sarkar, Md. Robiul Islam, Md. Khalid Syfullah, Md. Tohidul Islam, Md. Faysal Ahamed, Mominul Ahsan, and Julfikar Haider. 2023. Multi-Scale CNN: An Explainable AI-Integrated Unique Deep Learning Framework for Lung-Affected Disease Classification. *Technologies* (2023). <https://api.semanticscholar.org/CorpusID:263639725>
- [109] Yoshihide Sawada and Keigo Nakamura. 2022. Concept Bottleneck Model With Additional Unsupervised Concepts. *IEEE Access* 10 (2022), 41758–41765. <https://api.semanticscholar.org/CorpusID:246485723>
- [110] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128 (2020), 336–359. DOI: <http://dx.doi.org/10.1007/s11263-019-01228-7>
- [111] Andrei Semenov, Vladimir Ivanov, Aleksandr Beznosikov, and Alexander V. Gashnikov. 2024. Sparse Concept Bottleneck Models: Gumbel Tricks in Contrastive Learning. *ArXiv abs/2404.03323* (2024). <https://api.semanticscholar.org/CorpusID:268889881>
- [112] Md Shajalal, Sebastian Denef, Md. Rezaul Karim, Alexander Boden, and Gunnar Stevens. 2023. Unveiling Black-Boxes: Explainable Deep Learning Models for Patent Classification. *ArXiv abs/2310.20478* (2023). <https://api.semanticscholar.org/CorpusID:264492677>
- [113] Sharath M. Shankaranarayana and Davor Runje. 2019. ALIME: Autoencoder Based Approach for Local Interpretability. In *Ideal*. <https://api.semanticscholar.org/CorpusID:202539758>

- [114] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* abs/1312.6034 (2013). <https://api.semanticscholar.org/CorpusID:1450294>
- [115] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). <https://api.semanticscholar.org/CorpusID:14124313>
- [116] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. 2024. BioCLIP: A Vision Foundation Model for the Tree of Life. (2024). <https://arxiv.org/abs/2311.18803>
- [117] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. (2020). <https://arxiv.org/abs/1908.08530>
- [118] Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. 2023. Explain Any Concept: Segment Anything Meets Concept-Based Explanation. *ArXiv* abs/2305.10289 (2023). <https://api.semanticscholar.org/CorpusID:258740705>
- [119] Chung-En Sun, Tuomas P. Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2024. Concept Bottleneck Large Language Models. *ArXiv* abs/2412.07992 (2024). <https://api.semanticscholar.org/CorpusID:274638138>
- [120] Zeren Tan, Yang Tian, and Jian Li. 2023. GLIME: General, Stable and Local LIME Explanation. *ArXiv* abs/2311.15722 (2023). <https://api.semanticscholar.org/CorpusID:265455929>
- [121] Erico Tjoa and Cuntai Guan. 2022. Quantifying Explainability of Saliency Methods in Deep Neural Networks with a Synthetic Dataset. (2022). <https://arxiv.org/abs/2009.02899>
- [122] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv* abs/1807.03748 (2018). <https://api.semanticscholar.org/CorpusID:49670925>
- [123] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. DOI: <http://dx.doi.org/10.48550/arXiv.1706.03762>
- [124] Johanna Vielhaben, Stefan Blücher, and Nils Strodthoff. 2023. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Trans. Mach. Learn. Res.* 2023 (2023). <https://api.semanticscholar.org/CorpusID:256358599>
- [125] Patrick Wagner, Temesgen Mehari, Wilhelm Haverkamp, and Nils Strodthoff. 2023. Explaining Deep Learning for ECG Analysis: Building Blocks for Auditing and Knowledge Discovery. *Computers in biology and medicine* 176 (2023), 108525. <https://api.semanticscholar.org/CorpusID:258947616>

- [126] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3462–3471. DOI:<http://dx.doi.org/10.1109/cvpr.2017.369>
- [127] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *International Conference on Learning Representations*. [https://openreview.net/forum?id=GUrhfTuf\\_3](https://openreview.net/forum?id=GUrhfTuf_3)
- [128] S. Weber, M. Wyszynski, M. Godefroid, R. Plattfaut, and B. Niehaves. 2024. How do medical professionals make sense (or not) of AI? A social-media-based computational grounded theory study and an online survey. *Computational and Structural Biotechnology Journal* 24 (Feb 2024), 146–159. DOI:<http://dx.doi.org/10.1016/j.csbj.2024.02.009>
- [129] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. 2024. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. (2024). <https://arxiv.org/abs/2312.08344>
- [130] Xu Xiang, Hong Yu, Ye Wang, and Guoyin Wang. 2023. Stable local interpretable model-agnostic explanations based on a variational autoencoder. *Applied Intelligence* 53 (2023), 28226 – 28240. <https://api.semanticscholar.org/CorpusID:263094516>
- [131] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9653–9663. <https://arxiv.org/abs/2111.09886>
- [132] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, and Julian McAuley. 2023. Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models. (2023). <https://arxiv.org/abs/2310.03182>
- [133] Karina Yang, Alexis Bennett, and Dominique Duncan. 2023. Robust and Interpretable COVID-19 Diagnosis on Chest X-ray Images using Adversarial Training. *ArXiv abs/2311.14227* (2023). <https://api.semanticscholar.org/CorpusID:265445691>
- [134] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2022. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 19187–19197. <https://api.semanticscholar.org/CorpusID:253735286>
- [135] Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. *ArXiv abs/2111.07783* (2021). <https://api.semanticscholar.org/CorpusID:244117525>

- [136] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems* 33 (2020), 20554–20565. <https://arxiv.org/abs/1910.07969>
- [137] Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc Concept Bottleneck Models. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2205.15480> Spotlight Paper.
- [138] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moungh-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. 2024. Biomed-Parser: a biomedical foundation model for image parsing of everything everywhere all at once. (2024). <https://arxiv.org/abs/2405.12971>